THE FUTURE OF FREE SPEECH

THAT VIOLATES MY POLICIES

AI LAWS, CHATBOTS, AND THE FUTURE OF EXPRESSION

Directed by
Jordi Calvet-Bademunt,
Jacob Mchangama, and Isabelle Anzabi

OCTOBER 2025

Acknowledgments

The Future of Free Speech is an independent, nonpartisan think tank based at Vanderbilt University. Our mission is to reaffirm freedom of expression as the foundation of free and thriving societies through actionable research, practical tools, and principled advocacy. We envision a world in which the right to freedom of expression is safeguarded by law and strengthened by a culture that embraces diverse viewpoints.

This project was led by Jordi Calvet-Bademunt (Senior Research Fellow), Jacob Mchangama (Executive Director), and Isabelle Anzabi (Research Associate) at The Future of Free Speech. Together, they also drafted the chapters on the European Union and the United States of America.

We are grateful to Justin Hayes, Director of Communications, for overseeing the design of the report; Wendy H. Burch, Chief Operating Officer, for coordinating all administrative aspects of the project; and Sam Cosby, Director of Development, for leading the funding efforts that made this work possible.

We extend our thanks to the leading experts who contributed chapters on their respective jurisdictions: Carlos Affonso Souza (Brazil), Ge Chen (China), Sangeeta Mahapatra (India), and Kyung Sin (K.S.) Park (Republic of Korea). We are also grateful to Kevin T. Greene and Jacob N. Shapiro of Princeton University for their chapter, "Measuring Free Expression in Generative AI Tools."

We thank all the experts who contributed to individual chapters of this report; their names are listed in the relevant sections.

We are further indebted to Barbie Halaby of Monocle Editing for her careful editorial work across all chapters, and to Design Pickle for the report's design.

Finally, we are especially grateful to the Rising Tide Foundation and the Swedish Postcode Lottery Foundation for their generous support of this work, and we thank Vanderbilt University for their collaboration with and support of The Future of Free Speech.

© The Future of Free Speech 2025
Publication Date: October 2025

Preface

In this report, we explore the ways in which public and private governance of generative artificial intelligence (AI) shape the space for free expression and access to information in the 21st century.

Since the launch of ChatGPT by OpenAI in November 2022, generative AI has captured the public imagination. In less than three years, hundreds of millions of people have adopted OpenAI's chatbot and similar tools for learning, entertainment, and work.[1] Anthropic, another AI giant, now serves more than 300,000 business customers.[2] AI companies are valued in the hundreds of billions of US dollars[3], while established technology giants such as Google, Meta, and Microsoft are investing billions in the race to dominate the field.[4]

Generative AI refers to systems that create content — including text, images, video, audio, and software code — in response to user prompts.[5] Chatbots such as ChatGPT are the most visible examples, but generative AI is rapidly being embedded into the tools people use every day for both communication and access to information, from social media and email to word processors and search engines.

Recognizing generative AI's potential for expression and access to information, The Future of Free Speech undertook a first-of-its-kind analysis of freedom of expression in major models. In February 2024, we assessed the "free-speech culture" of six leading systems, focusing on their usage policies and responses to prompts.[6] Our findings revealed that excessively broad and vague rules often resulted in undue restrictions on speech and access to information.[7] By April 2025, when we updated this work, we observed signs of change: Some models showed greater openness.[8]

This current report builds on those foundations and pursues a more ambitious goal. Supported by leading experts, The Future of Free Speech undertakes a deeper examination of how national legislation and corporate practices shape freedom of expression in the era of generative AI. "*That Violates My Policies": AI Laws, Chatbots, and the Future of Expression* explores:

- AI legislation in Brazil, China, the European Union, India, the Republic of Korea, and the United States.[9] In this report, AI legislation refers to laws and public policies addressing AI-generated content, with particular focus on elections and political speech, hate speech, defamation, explicit content (including

1 MacKenzie Sigalos, "OpenAI's ChatGPT to Hit 700 Million Weekly Users, Up 4x from Last Year," *CNBC*, August 4, 2025, https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html.
2 Hayden Field, "Anthropic Is Now Valued at $183 Billion," *The Verge*, September 2, 2025, https://www.theverge.com/anthropic/769179/anthropic-is-now-valued-at-183-billion.
3 Kylie Robison, "OpenAI Is Poised to Become the Most Valuable Startup Ever: Should It Be?," *Wired*, August 19, 2025, https://www.wired.com/story/openai-valuation-500-billion-skepticism/; Krystal Hu and Shivani Tanna, "OpenAI Eyes $500 Billion Valuation in Potential Employee Share Sale, Source Says," *Reuters*, August 6, 2025, https://www.reuters.com/business/openai-eyes-500-billion-valuation-potential-employee-share-sale-source-says-2025-08-06/.
4 Blake Montgomery, "Big Tech Has Spent $155bn on AI This Year: It's About to Spend Hundreds of Billions More," *The Guardian*, August 2, 2025, https://www.theguardian.com/technology/2025/aug/02/big-tech-ai-spending.
5 Cole Stryker and Mark Scapicchio, "What Is Generative AI?," IBM Think, March 22, 2024, https://www.ibm.com/think/topics/generative-ai.
6 Jordi Calvet-Bademunt and Jacob Mchangama, *Freedom of Expression in Generative AI: A Snapshot of Content Policies* (Future of Free Speech, February 2024), https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.
7 Calvet-Bademunt and Mchangama, *Freedom of Expression in Generative AI*.
8 Jordi Calvet-Bademunt, Jacob Mchangama, and Isabelle Anzabi, "One Year Later: AI Chatbots Show Progress on Free Speech — But Some Concerns Remain," *The Bedrock Principle*, April 1, 2025, https://www.bedrockprinciple.com/p/one-year-later-ai-chatbots-show-progress.
9 To select the countries, we considered Stanford University's 2023 Global AI Vibrancy Ranking (the most recent available at the time of writing), along with factors such as geographic diversity, population size, democratic and freedom status, and the presence of existing or emerging AI-related legislation.

child sexual abuse material and nonconsensual intimate images), and copyright. We also consider measures that actively promote freedom of expression, such as AI literacy initiatives and policies supporting cultural and linguistic diversity.

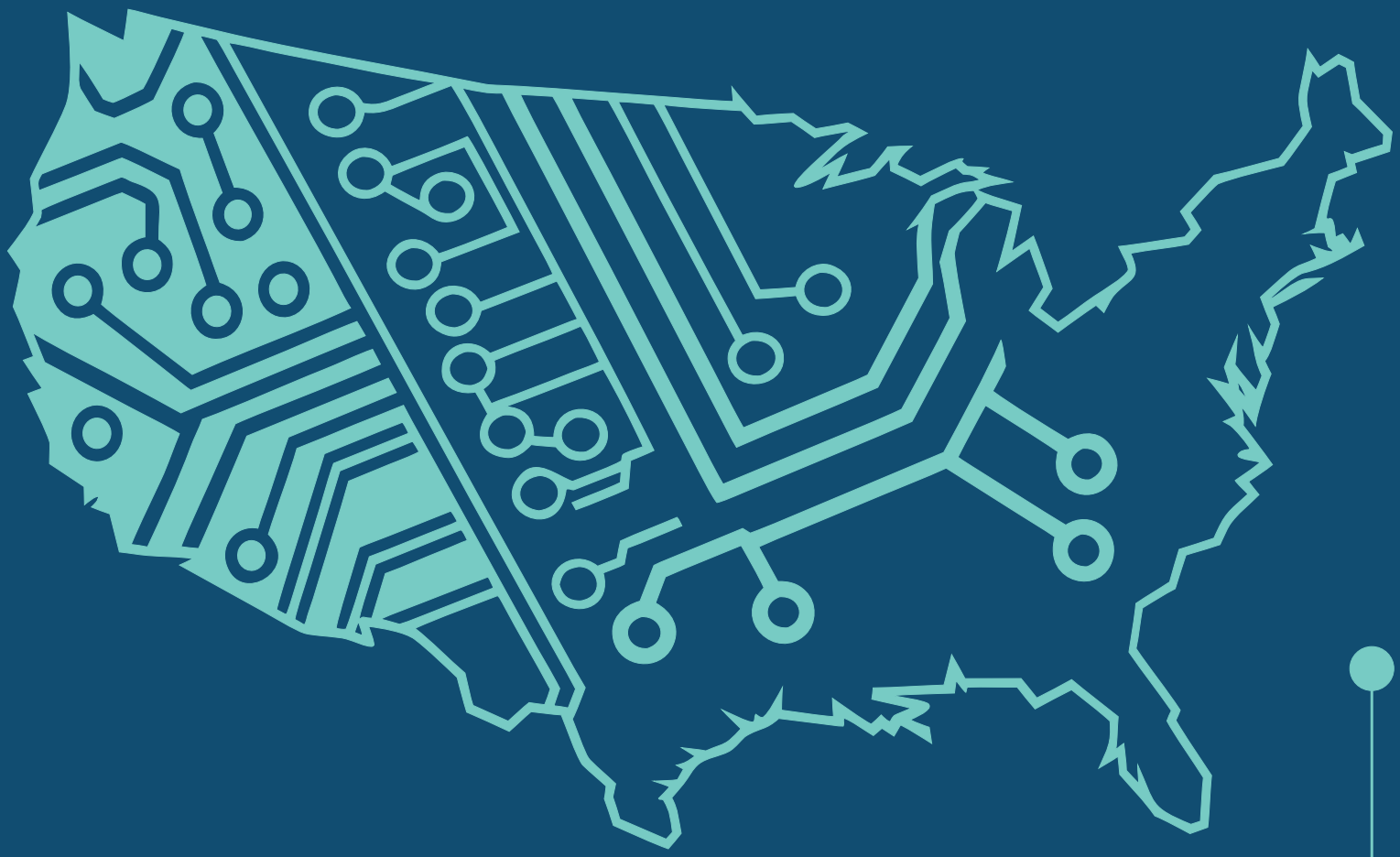
- Corporate practices of major AI developers, including Alibaba, Anthropic, Google, Meta, Mistral AI, DeepSeek, OpenAI, and xAI.[10] We examine their usage policies, model performance in responding to prompts, and the limited available information on their training data and development processes.

This report seeks to provide a rigorous and timely analysis of how generative AI is reshaping the space for free expression in both the public and private spheres. Building on these insights, The Future of Free Speech is developing guidelines to help policymakers and companies ensure that generative AI protects and enhances freedom of expression and access to information, two cornerstones of democratic societies.

In an era of rapid technological change, safeguarding free expression is a matter not only of rights but of preserving the conditions for open, informed, and thriving democracies. developing guidelines to help policymakers and companies ensure that generative AI protects and enhances freedom of expression and access to information, two cornerstones of democratic societies.

In an era of rapid technological change, safeguarding free expression is a matter not only of rights but of preserving the conditions for open, informed, and thriving democracies.

10 We selected major models from leading companies that are accessible through a web interface and include text-generation capabilities. In addition, we considered the geographic location of the model provider and the degree of openness of the models.



Artificial Intelligence and Freedom of Expression in the United States

Isabelle Anzabi, Jordi Calvet-Bademunt, and Jacob Mchangama*

* Isabelle Anzabi, Jordi Calvet-Bademunt, and Jacob Mchangama are a research associate, senior research fellow, and executive director, respectively, at the Future of Free Speech. We thank Jeff Kosseff and Ashkhen Kazaryan for their valuable comments and suggestions. All remaining errors are our own.

Abstract

The rapid rise of generative artificial intelligence (AI) in the United States is testing the resilience of established free speech protections. This chapter examines the evolving legal and policy landscape as lawmakers, courts, and regulators confront how AI intersects with defamation, political expression, copyright, and other core areas of speech law. The lack of federal AI regulation has prompted a patchwork of state measures on issues such as political deepfakes, disclosure mandates, algorithmic discrimination, and explicit content. These developments have intensified debates over liability for AI-generated harms and the proper scope of regulation without eroding First Amendment guarantees. While the US approach affords a high degree of expressive freedom compared with many jurisdictions, it is marked by a heavy reliance on judicial interpretation to resolve novel disputes. As AI-generated speech becomes increasingly important, we underscore that any regulatory response must remain tightly focused on preventing real, direct, and imminent harms — to ensure constitutional principles are preserved and the free exchange of ideas remains a defining feature of the American legal order.

Isabelle Anzabi

Isabelle Anzabi is a research associate at The Future of Free Speech, where she analyzes the intersections between AI policy and freedom of expression. She is bringing her background in digital rights policy and global regulatory approaches to content moderation and AI governance. Previously, Isabelle was an AI & Human Rights Fellow with the European Center for Not-for-Profit Law, a Knowledge Fellow at the DiploFoundation, and a research group member at the Center for AI and Digital Policy. Isabelle received her B.A. in Political Science from Stanford University. She also studied digital governance at Oxford University and interned at institutions such as the World Bank and CISA. On campus, Isabelle was affiliated with the Stanford Center for Racial Justice, the Stanford Legal Design Lab, the Stanford Cyber Policy Center, the Stanford Constitutional Law Center, the Stanford Technology Law Review, and the Public Service Leadership Program.

Jordi Calvet-Bademunt

Jordi Calvet-Bademunt is a Senior Research Fellow at The Future of Free Speech. He is also a Visiting Legal Researcher at the Barcelona Supercomputing Center, where he advises on trustworthy AI. His work focuses on AI policy and digital governance, and he has written extensively and provided commentary in both specialist and mainstream media. Previously, Jordi spent about a decade working at the Organisation for Economic Co-operation and Development (OECD) and as an associate at leading European law firms. He holds advanced degrees from Harvard University and the College of Europe in Bruges, Belgium.

Jacob Mchangama

Jacob Mchangama is the Founder and Executive Director of The Future of Free Speech. He is a research professor at Vanderbilt University and a Senior Fellow at The Foundation for Individual Rights and Expression (FIRE). In 2018, he was a visiting scholar at Columbia's Global Freedom of Expression Center. He has commented extensively on free speech and human rights in outlets including the Washington Post, the Wall Street Journal, The Economist, Foreign Affairs and Foreign Policy. Jacob has published in academic and peer-reviewed journals, including Human Rights Quarterly, Policy Review, and Amnesty International's Strategic Studies. He is the producer and narrator of the podcast "Clear and Present" Danger: A History of Free Speech and the critically acclaimed book Free Speech: A History From Socrates to Social Media, published by Basic Books in 2022. He is the recipient of numerous awards for his work on free speech and human rights.

1. Introduction

The United States has one of the most robust systems of free speech protection in the world, anchored in the First Amendment command that "Congress shall make no law … abridging the freedom of speech, or of the press."[1] The Supreme Court has interpreted this protection broadly, extending it to new communication technologies and safeguarding both the right to speak and the right to receive information and ideas. Generative AI presents the next major test of these principles.

Generative AI's capacity to produce text, images, audio, and video at scale offers unprecedented opportunities to expand access to information, amplify diverse voices, and lower barriers to participation in public debate. It can serve as a creative and educational tool, a means of preserving cultural and linguistic diversity, and a way to make information more accessible to people with different needs and backgrounds. Yet the same capabilities raise novel questions about liability, truthfulness, and the potential for misuse — from defamation and political deepfakes to the unauthorized reproduction of copyrighted works.

The rapid pace of AI development has outstripped the adoption of comprehensive federal legislation, leaving a fragmented legal environment in which states have taken varied approaches. These range from regulating algorithmic discrimination, disclosure requirements, and frontier model safety to addressing explicit content and political manipulation. Courts are beginning to confront whether and to what extent AI-generated content should receive First Amendment protection and how existing doctrines on defamation, third-party immunity, and copyright apply when the "speaker" may not be human.

In this chapter we explore these emerging challenges in detail, examining federal and state regulatory trends, the unsettled question of AI's status under the First Amendment, and the constitutional limits on regulating harmful or deceptive content. While acknowledging legitimate concerns about AI's potential misuse, we argue that the United States should address these risks in ways that preserve the country's long-standing commitment to protecting even controversial or offensive expression. In the age of generative AI, safeguarding the open exchange of ideas remains not only a constitutional imperative but also a prerequisite for ensuring that this transformative technology strengthens — rather than constrains — the freedom to speak and to know.

1 U.S. Const. amend. I.

2. Substantive Analyses

2.1. General Standards of Freedom of Expression

The cornerstone of free expression in the United States is the First Amendment. This foundational principle, ratified in 1791,[2] explicitly prohibits Congress from enacting any law that abridges freedom of speech and protects access to information and ideas.[3] Notably, the First Amendment protects the right to receive information and ideas "regardless of their social worth,"[4] underscoring the protection against state-imposed viewpoint discrimination.

The Supreme Court has established a framework wherein different categories of speech receive varying levels of protection under the First Amendment. Political, ideological, and artistic speech are considered at the core of the First Amendment.[5] While also a protected communication under the First Amendment, commercial speech receives less protection than other forms of speech.[6] Additionally, the court has identified specific categories of speech that may be regulated. These categories, unprotected by the First Amendment, include incitement to imminent lawless action, true threats, fraud, defamation, obscenity, and child pornography (also referred to as child sexual abuse material or CSAM).[7]

Determining the constitutionality of regulations of protected speech hinges on whether the regulation is content-based or content-neutral.[8] Content-based restrictions are not automatically unconstitutional, but they are subject to strict scrutiny, which is an exceptionally high standard that requires the government to demonstrate that the law is the least restrictive means of advancing a compelling governmental interest.[9] In contrast, content-neutral restrictions, such as time, place, and manner regulations, are evaluated under intermediate scrutiny, which requires the law to be narrowly tailored to serve a substantial governmental interest.[10] Commercial speech also receives intermediate scrutiny. The Supreme Court has held that viewpoint discrimination — a subset of content discrimination in which the government targets or favors specific opinions or beliefs — is the most "egregious."[11]

The court has historically adapted First Amendment principles to new technologies, recognizing that the right to free expression extends beyond traditional forms of communication to encompass novel innovations.[12]

2 "The Bill of Rights: A Transcription," National Archives, archived November 4, 2015, https://www.archives.gov/founding-docs/bill-of-rights-transcript.
3 Lamont v. Postmaster General, 381 U.S. 301 (1965)
4 Stanley v. Georgia, 394 U.S. 557 (1969).
5 Congress.gov, "The First Amendment: Categories of Speech," March 28, 2024, https://www.congress.gov/crs-product/IF11072.
6 Central Hudson Gas & Elec. v. Public Svc. Comm'n, 447 U.S. 557 (1980).
7 Congress.gov, "The First Amendment: Categories of Speech."
8 "A content-based law or regulation discriminates against speech based on the substance of what it communicates." David L. Hudson Jr., "Content Based," The Free Speech Center, August 10, 2023, https://firstamendment.mtsu.edu/article/content-based. "Content neutral refers to laws that apply to all expression without regard to the substance or message of the expression." David L. Hudson Jr., "Content Neutral," The Free Speech Center, January 1, 2009, https://firstamendment.mtsu.edu/article/content-neutral
9 Congress.gov, "Free Speech: When and Why Content-Based Laws Are Presumptively Unconstitutional," January 10, 2023, https://www.congress.gov/crs-product/IF12308.
10"Overview of Content-Based and Content-Neutral Regulation of Speech," Constitution Annotated, Library of Congress, accessed August 1, 2025, https://constitution.congress.gov/browse/essay/amdt1-7-3-1/ALDE_00013695/.
11 "Overview of Viewpoint-Based Regulation of Speech," Legal Information Institute, accessed August 1, 2025, https://www.law.cornell.edu/constitution-conan/amendment-1/overview-of-viewpoint-based-regulation-of-speech
12 Brown, et al. v. Entertainment Merchants Assn. et al., 564 U.S. 786 (2011).

In 1997, the court advanced its First Amendment jurisprudence to the internet by ruling that even well-intentioned government regulations can be struck down as overly broad.[13] This adaptability suggests that the core tenets of free speech will likely be afforded to AI-generated content as well. However, in 2025, the court limited some free speech protections by upholding a Texas law that required age verification for websites if one-third of the content is sexual material harmful to minors.[14]

Section 230 of the Communications Decency Act shapes the digital speech landscape by shielding internet service providers and platforms from liability for content not generated by the platform.[15] The possibility that Section 230 could apply to generative AI content raises complex legal questions about responsibility and accountability for speech not authored by a human.

The application of freedom of expression standards to content generated by AI is a subject of ongoing legal debate. A fundamental question is whether AI-generated content even constitutes "speech" under the First Amendment.

There are strong reasons to consider protecting AI-generated content under the First Amendment, with legal scholars arguing that the focus should be on the listener's right to receive information — regardless of the source being human or artificial.[16] The Supreme Court recognizes the right to receive information as a corollary to the right to speak, aligning with the perspective that users have a right to obtain information from AI models. Some scholars have suggested that even AI output generated with no human intervention should be protected.[17] Generative AI is a tool for creating expressive content, and similar to the press or cameras, it "make[s] it easier to speak."[18] Scholars have pointed out that the First Amendment protects the rights of creators and users, and restricting AI-generated content could infringe on their rights when using AI to express themselves.[19]

Still, some argue against full First Amendment protection for AI output, viewing it as not inherently expressive or as lacking the human intentionality that traditionally underlies free speech rights.[20] This view suggests that generative AI, particularly large language models (LLMs), may not be "speaking" in a way that warrants constitutional protection but are rather generating automated responses based on algorithms and training data. While litigation against Character Technologies over its chatbot Character.AI is still ongoing, US District Judge Anne Conway rejected some arguments that chatbots are protected by the First Amendment, stating "the Court is not prepared to hold that Character A.I.'s output is speech" at this stage of the litigation.[21] Moreover, Judge Conway asserted that "Defendants can assert the First Amendment rights of its users," who have the right to receive the speech of chatbots.[22] Ultimately though, this is a district court decision, not binding in other jurisdictions and subject to appeal.

13 Reno v. ACLU, 521 U.S. 844 (1997).
14 Free Speech Coalition, Inc. v. Paxton, 606 U.S. ___ (2025).
15 Congress.gov, "Section 230: An Overview," January 4, 2025, https://www.congress.gov/crs-product/R46751.
16 Jane R. Yakowitz Bambauer, "Negligent AI Speech: Some Thoughts About Duty," *Journal of Free Speech Law*, April 28, 2023, http://dx.doi.org/10.2139/ssrn.4432822.
17 Toni Marie Massaro, Helen L. Norton, and Margot E. Kaminski, "SIRI-OUSLY 2.0: What Artificial Intelligence Reveals about the First Amendment," *Minnesota Law Review* 101 (June 28, 2017): 2481, https://www.minnesotalawreview.org/wp-content/uploads/2019/07/MassaroNortonKaminski-1.pdf.
18 Volokh, Lemley, and Henderson, "Freedom of Speech and AI Output," 658.
19 Eugene Volokh, Mark A. Lemley, and Peter Henderson, "Freedom of Speech and AI Output," *Journal of Free Speech Law* 3 (August 3, 2023): 651, https://www.journaloffreespeechlaw.org/volokh lemleyhenderson.pdf; "AI and the First Amendment: A Q&A with Jack Balkin," Yale Law School, January 29, 2024, https://law.yale.edu/yls-today/news/ai-and-first-amendment-qa-jack-balkin.
20 Peter Salib, "AI Outputs Are Not Protected Speech," *Washington University Law Review* (forthcoming), University of Houston Law Center Research Paper no. 2024-A-5, January 1, 2024, https://ssrn.com/abstract=4687558; Karl M. Manheim and Jeffery Atik, "AI Outputs and the Limited Reach of the First Amendment," *Washburn Law Journal* 63 (2024): 159, https://ssrn.com/abstract=4626235.
21 Reply in Support of Motion to Dismiss, Garcia v. Character Technologies, Inc., No. 6:24-cv-01903-ACC-UAM (M.D. Fla. May 21, 2025), ECF No. 115, 31, https://storage.courtlistener.com/recap/gov.uscourts.flmd.433581/gov.uscourts.flmd.433581.115.0.pdf; Kate Payne, "In Lawsuit over Teen's Death, Judge Rejects Arguments That AI Chatbots Have Free Speech Rights," AP News, May 21, 2025, https://apnews.com/article/ai-lawsuit-suicide-artificial-intelligence-free-speech-ccc77a5ff5a84bda753d2b044c83d4b6; Adi Robertson, "Are Character AI's Chatbots Protected Speech? One Court Isn't Sure," The Verge, May 21, 2025, https://www.theverge.com/law/672209/character-ai-lawsuit-ruling-first-amendment.
22 Reply in Support of Motion to Dismiss, 27.

The First Amendment generally protects the publication of AI-generated content by users, subject to the same restrictions as human speech. Distributing AI content is no different than distributing information or opinions obtained from any other source. This means users are protected from government intervention when sharing AI content but could face liability for the content they publish. For example, if AI is used to generate defamatory content, the user who publishes that content could still be held liable under defamation laws, provided that the plaintiff overcomes the First Amendment protections afforded to defamation defendants. Similarly, AI could potentially generate content that incites violence or constitutes a true threat, which would also fall outside the scope of First Amendment protection.

2.2. AI-Specific Legislation and Policies

2.2.1. International Agreements

At an international level, the United States signed the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law in 2024.[23] This convention is the first binding international treaty on AI. It applies to activities within the life cycle of AI systems undertaken by public authorities or private actors acting on their behalf. Parties to the convention must address risks and impacts arising from private actors, but they have flexibility in how to do so. The convention sets out seven core principles, including human dignity and individual autonomy, transparency, accountability, and privacy. It also establishes obligations to protect human rights, safeguard the integrity of democratic processes, and uphold respect for the rule of law.

2.2.2. Federal Efforts

As of August 2025, the United States has not adopted a comprehensive federal framework for the regulation of AI. Instead, AI policy has relied mainly on initiatives from the executive branch. Federal policy has undergone a marked transformation, reflecting a deliberate pivot toward deregulation and innovation. This shift was formalized on the first day of President Donald Trump's second administration through the issuance of the executive order titled Initial Rescissions of Harmful Executive Orders and Actions,[24] which revoked President Joe Biden's 2023 executive order Safe, Secure, and Trustworthy Artificial Intelligence.[25] The rescission signaled a decisive departure from the previous administration's precautionary approach, which emphasized civil rights protections, algorithmic oversight, and risk management. The Trump administration is promoting use of open AI models with the issuance of the executive order Removing Barriers to American Leadership in Artificial Intelligence, which articulates a deregulatory philosophy rooted in global competitiveness and national sovereignty.[26] The order asserts that American AI development must be "free from ideological bias or engineered social agendas" and called for the creation of a national AI Action Plan.

This approach has been operationalized through agency guidance. In April 2025, the Office of Management and Budget (OMB) issued two complementary memoranda, M-25-21 and M-25-22,[27] offering updated

23 Council of Europe: Committee of Ministers, "Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law," CETS No. 25, May 17, 2024, https://rm.coe.int/1680afae3c
24 Exec. Order No. 14148, 90 Fed. Reg. 8237 (January 20, 2025), https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-action
25 Exec. Order No. 14148; Exec. Order No. 14110, 88 Fed. Reg. 75191 (November 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
26 Exec. Order No. 14179, 90 Fed. Reg. 8741 (January 31, 2025), https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence.
27 OMB Memorandum M-25-21, "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust," April 3, 2025, https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf; and OMB Memorandum M-25-22, "Driving Efficient Acquisition of Artificial Intelligence in Government," April 3, 2025, https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf.

directives on AI procurement, risk classification, and oversight across the executive branch.[28] The new guidance is intended to simplify internal compliance procedures while ensuring that systems with potential implications for civil liberties or public safety are subject to heightened scrutiny.[29]

In July 2025, the Trump administration formalized its deregulatory posture on AI governance through the release of America's AI Action Plan[30] and a new executive order, Preventing Woke AI in the Federal Government.[31] These directives frame federal procurement as a tool for promoting what the administration considers "objective" AI systems — those free from perceived ideological influence — and prioritizing open-source and open-weight models to ensure transparency and prevent centralized control over AI capabilities. As part of the plan's implementation, the National Institute of Standards and Technology (NIST) was directed to revise its widely adopted AI Risk Management Framework to remove all references to misinformation, diversity, equity, and inclusion (DEI), and climate change — terms that have become flash points in US debates over free expression. The plan also instructs NIST's Center for AI Standards and Innovation to evaluate frontier AI models developed in the People's Republic of China for alignment with Chinese Communist Party propaganda. Although framed as a pushback against foreign censorship, this directive raises questions about the limits of viewpoint neutrality in federal AI policy.[32]

The White House's AI Action Plan advocates for the development and use of open-source and open-weight AI models.[33] The plan promotes a supportive environment for open models, with a focus on investment and streamlined access to computing resources. It frames open development as a way to enhance transparency, accelerate innovation, and set global standards. Open models offer a counterbalance to centralized control, enabling diverse communities to shape systems according to their own values and needs.[34]

2.2.3. State-Level Efforts

The proliferation of AI has prompted an assertive legislative response at the state level in the United States. In the absence of a unified federal AI framework, states have emerged as primary actors in shaping the legal and normative contours of AI governance. During the 2025 legislative session, all 50 states, Puerto Rico, the Virgin Islands, and Washington, DC, introduced AI-related bills, with 38 states having adopted or enacted approximately 100 measures.[35] While these legislative experiments underscore states' roles as laboratories of democracy, they also raise profound questions about federalism, preemption, and the constitutional limits of state power, particularly under the First Amendment.

Congress attempted to pass a 10-year moratorium on state-level AI enforcement, which ultimately failed. The House of Representatives passed a version along partisan lines that stated "no State or political subdivision thereof may enforce any law or regulation regulating artificial intelligence models, artificial intelligence systems,

28 "Fact Sheet: Eliminating Barriers for Federal Artificial Intelligence Use and Procurement," The White House, April 7, 2025, https://www.whitehouse.gov/fact-sheets/2025/04/fact-sheet-eliminating-barriers-for-federal-artificial-intelligence-use-and-procurement; "White House Releases New Policies on Federal Agency AI Use and Procurement," The White House, April 7, 2025, https://www.whitehouse.gov/articles/2025/04/white-house-releases-new-policies-on-federal-agency-ai-use-and-procurement

29 "Fact Sheet: Eliminating Barriers for Federal AI Use."

30 "White House Unveils America's AI Action Plan," The White House, July 23, 2025, https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan

31 Exec. Order No. 14319, 90 Fed. Reg. 35389 (July 23, 2025), https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government.

32 Isabelle Anzabi and Jordi Calvet-Bademunt, "The Anti-'Woke' AI Agenda & Free Speech," *The Bedrock Principle*, July 23, 2025, https://www.bedrockprinciple.com/p/the-anti-woke-ai-agenda-free-speech.

33 "White House Unveils America's AI Action Plan."

34 Isabelle Anzabi, "The Future of Free Speech's Comments on the U.S. AI Action Plan," *The Bedrock Principle*, March 24, 2025, https://www.bedrockprinciple.com/p/the-future-of-free-speechs-comments.

35 National Conference of State Legislatures, "Artificial Intelligence 2025 Legislation," NCSL, accessed August 1, 2025, https://www.ncsl.org/technology-and-communication/artificial-intelligence-2025-legislation.

or automated decisions during the 10-year period" following its enactment.[36] However, following revisions, the US Senate voted to strike it.[37]

States are enacting generative AI regulations, targeting six core concerns: high-risk AI systems and algorithmic discrimination; disclosure and labeling requirements; frontier model safety; access to computation and accountability; explicit content, covered in section 2.4; and political deepfakes and deceptive media, which we address in section 2.6.

2.2.3.1. High-Risk AI Systems and Algorithmic Discrimination

One of the most prominent state initiatives is the Colorado Artificial Intelligence Act (CAIA), which establishes a regulatory framework for "high-risk" AI systems, defined as those that significantly affect individuals' legal rights or access to essential services.[38] Although the law's implementation date has been delayed, CAIA imposes a duty of care on developers and deployers to prevent algorithmic discrimination, and it mandates transparency mechanisms such as consumer notices and annual impact assessments.[39] Importantly, CAIA exempts chatbots that communicate with "consumers in natural language for the purpose of providing users with information" and that are "subject to an accepted use policy that prohibits generating content that is discriminatory or harmful." The accepted use policy is not detailed further and does not define "harmful." This raises concerns as the law effectively requires the implementation of content restrictions, which may compel private actors to adopt policies that restrict protected speech categories to avoid liability.

Enacted in June 2025, the Texas Responsible Artificial Intelligence Governance Act (TRAIGA) prohibits intentionally developing and deploying AI systems for behavioral manipulation (encouraging physical harm or criminal activity), constitutional infringement (restricting federal constitutional rights), unlawful discrimination (targeting protected classes), and harmful content creation (producing CSAM, unlawful deepfakes, or explicit content involving minors).[40] TRAIGA explicitly states, "This chapter may not be construed to: (1) impose a requirement on a person that adversely affects the rights or freedoms of any person, including the right of free speech." The revised version departs from earlier drafts criticized for their broad innovation-stifling mandates[41] and for including a provision prohibiting AI systems from engaging in "political viewpoint discrimination."[42]

Virginia's now-vetoed High-Risk Artificial Intelligence Developer and Deployer Act (HB 2094) would have imposed similar obligations on developers of high-risk AI systems to document system limitations, ensure transparency, and manage risks associated with algorithmic discrimination. Additionally, deployers would have had to disclose AI usage to consumers and conduct impact assessments. However, Governor Glenn Youngkin vetoed the bill in March 2025, citing existing laws and concerns of overburdening small businesses and stifling innovation.[43]

---

36 One Big Beautiful Bill Act, H.R. 1, 119th Cong. (2025–2026), § 43201(c), May 22, 2025, https://www.congress.gov/bill/119th-congress/house-bill/1/text/eh.
37 Billy Perrigo and Andrew R. Chow, "Senators Reject 10-Year Ban on State-Level AI Regulation in Blow to Big Tech," *Time*, July 1, 2025, https://time.com/7299044/senators-reject-10-year-ban-on-state-level-ai-regulation-in-blow-to-big-tech
38 S.B. 205, "Concerning Consumer Protections in Interactions with Artificial Intelligence Systems," 2024 Reg. Sess. (Colo. 2024), enacted May 17, 2024, https://leg.colorado.gov/bills/sb24-205.
39 "Colorado Passes Bill Amending Current AI Legislation, " *GovTech*, September 3, 2025, https://www.govtech.com/artificial-intelligence/colorado-passes-bill-amending-current-ai-legislation
40 H.B. 149, "Texas Responsible Artificial Intelligence Governance Act," 89th Leg., Reg. Sess. (Tex. 2025) (enacted June 22, 2025; effective Jan. 1, 2026), https://capitol.texas.gov/tlodocs/89R/billtext/pdf/HB00149F.pdf; Jason M. Loring and Graham H. Ryan, "Texas Enacts AI Law Targeting Harmful Use, Fostering Innovation," *National Law Review*, June 24, 2025, https://natlawreview.com/article/texas-enacts-responsible-ai-governance-act.
41 H.B. 1709, "Texas Responsible Artificial Intelligence Governance Act," 89th Leg., Reg. Sess. (Tex. 2025) (introduced version), https://capitol.texas.gov/tlodocs/89R/billtext/pdf/HB01709I.pdf#navpanes=0.
42 Austin Jenkins, "Capriglione Introduces Overhauled AI Bill in Texas," *Pluribus News*, March 18, 2025, https://pluribusnews.com/news-and-events/capriglione-introduces-overhauled-ai-bill-in-texas.
43 H.B. 2094, "High-Risk Artificial Intelligence; Definitions, Development, Deployment, and Use; Civil Penalties," 2025 Reg. Sess. (Va. 2025) (vetoed by Governor Mar. 24, 2025; House sustained veto Apr. 2, 2025), https://lis.virginia.gov/bill-details/20251/HB2094.

2.2.3.2. Disclosure and Labeling Requirements

Utah's AI Policy Act (UAIP), enacted in early 2024, requires generative AI disclosures in regulated professional contexts such as health care.[44] While the UAIP avoids many of the constitutional pitfalls that accompany broader compelled-disclosure regimes, even these targeted requirements may encounter First Amendment challenges if applied to expressive interactions in counseling, education, or other advisory contexts. These constitutional barriers are in place to protect against government overreach that could compel speech or interfere with free expression, a core principle of the First Amendment. Disclosure mandates may force developers, platforms, or creators to convey messages they do not endorse or alter the expressive intent of AI-generated content.

California's AI Transparency Act (SB 942) mandates the inclusion of both visible and invisible watermarks in AI-generated media.[45] Though such transparency mandates are aimed at combating misinformation and synthetic disinformation, their breadth and enforcement mechanisms both raise potential First Amendment issues, especially if they require speech by platforms or developers that conflicts with their editorial discretion or artistic intent. In California, the Training Data Transparency Act (AB 2013) requires developers to disclose information about the datasets used to train generative AI models.[46] Virginia's HB 2094 would have also mandated disclosure and labeling of synthetic content as a tool for mitigating misinformation, exemplifying the trend among states in this regard.[47]

Compelled disclosures involving expressive content — especially when broadly framed — risk being struck down as impermissible compelled speech under the First Amendment. Courts have long distinguished between commercial speech and expressive speech, and though the former may be subject to certain mandatory disclosures (e.g., in advertising or professional conduct), the latter is more robustly protected against government-imposed messaging. Thus, any legislative requirement that effectively mandates disclaimers on expressive AI outputs, such as political satire or artistic works, must undergo exacting constitutional scrutiny.

2.2.3.3. AI Safety and Frontier Model Regulation

California and New York have grappled with the constitutional and policy challenges of regulating frontier AI models, the most powerful and resource-intensive AI systems. California's attempt, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB 1047), was vetoed by Governor Gavin Newsom in September 2024.[48] The bill would have imposed a "duty of reasonable care" on developers to prevent "critical harm" and required a "kill switch" for models posing severe risks.[49] Governor Newsom's veto cited concerns that the bill's broad scope would stifle innovation and disproportionately burden smaller companies.

New York's Responsible AI Safety and Education (RAISE) Act (S6953B/A6453B), awaiting the governor's signature, takes a more targeted approach.[50] It applies only to the largest AI developers and focuses on

44 S. 149, "Artificial Intelligence Amendments," 2024 Gen. Sess. (Utah 2024) (enacted Mar. 13, 2024; effective May 1, 2024), https://le.utah.gov/~2024/bills/static/SB0149.html.
45 S.B. 942, "California AI Transparency Act," 2023–24 Reg. Sess. (Cal. 2024) (signed Sept. 19, 2024; chap. 291), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942.
46 A.B. 2013, "An Act to Add Title 15.2," 2023–24 Reg. Sess. (Cal. 2024) (approved by Governor Sept. 28, 2024), https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=202320240AB2013.
47 H.B. 2094, "High-Risk Artificial Intelligence" (Va. 2025)
48 Bobby Allyn, "California Gov. Newsom Vetoes AI Safety Bill That Divided Silicon Valley," *NPR*, September 29, 2024, https://www.npr.org/2024/09/20/nx-s1-5119792/newsom-ai-bill-california-sb1047-tech.
49 S.B. 1047, "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act," 2023–24 Reg. Sess. (Cal. 2024) (vetoed by Governor Sept. 29, 2024), https://legiscan.com/CA/text/SB1047/id/2919384.
50 S. 6953-B, "Responsible AI Safety and Education Act" (RAISE Act), 2025 Reg. Sess. (N.Y. 2025), https://www.nysenate.gov/legislation/bills/2025/S6953/amendment/B.

preventing the most severe risks, such as assisting in the creation of biological weapons.[51] The bill mandates safety plans and risk evaluations but avoids a "kill switch" requirement. While these frontier model regulations primarily concern physical and cyber safety, they have prompted debate over whether overly broad mandates could indirectly chill the development of models capable of generating a wide range of expressive content, thereby potentially impacting the innovation that underpins new forms of speech.

2.2.3.4. Access to Computation and Accountability

Emerging legislative models suggest a conceptual shift in how states view computational access as a right. Montana's Right to Compute Act (SB 212) frames access to AI and computation as a positive right, potentially inviting future litigation over whether restrictions on AI tools might infringe on constitutional or quasi-constitutional interests, such as freedom of expression or access to information.[52] Similarly, California's SB 53 on whistleblower protections for employees of foundational model developers reflects a growing emphasis on procedural safeguards and transparency within AI development and on the values that align with democratic accountability and public oversight.[53]

These varied state efforts illustrate the dynamic and experimental nature of state-level AI governance. They also expose a constitutional fault line: the risk that well-meaning regulation of AI systems inadvertently infringes on protected expressive conduct. As generative AI continues to serve as both a subject and a medium of speech, courts will increasingly be called upon to determine the permissible bounds of government intervention.

2.3. Defamation

2.3.1. Traditional Rules of Defamation and AI-Generated Content

The legal framework governing liability for AI-generated content remains unsettled, particularly in the absence of comprehensive federal legislation. In the current landscape, traditional doctrines of defamation, fraud, and intellectual property infringement are being adapted to address the unique challenges posed by AI systems. Central to this inquiry is a question: Who may be held legally responsible when an AI system produces harmful or unlawful speech?

Under established defamation principles, liability arises when a person "publishes" a false statement of fact about another that causes reputational harm. There must be some level of fault, which varies by state law. For statements about public figures, the plaintiff must also demonstrate actual malice — that the speaker knew the statement was false or acted with reckless disregard for the truth. While these rules were crafted in the context of human speakers, they are understood to extend to situations where a person uses a tool, such as an AI model, to create or disseminate defamatory content. Thus, a user who knowingly prompts an AI system to generate and then publicly shares a false and injurious statement could be liable under conventional defamation theory. Under the negligence standard, which typically applies to statements about private figures, a user who unknowingly publishes defamatory content may still be held liable for failing to exercise reasonable

51 Jennifer Johnson et al., "New York Legislature Passes Sweeping AI Safety Legislation," *Global Policy Watch*, June 24, 2025, https://www. globalpolicywatch.com/2025/06/new-york-legislature-passes-sweeping-ai-safety-legislation.
52 S. 212, "Creating the Right to Compute Act and Requiring Shutdowns of AI-Controlled Critical Infrastructure," 2025 Reg. Sess. (Mont. 2025) (signed by Governor Apr. 16, 2025; chapter assigned Apr. 17, 2025), https://bills.legmt.gov/#/laws/bill/2/LC0292
53 S.B. 53, "Artificial Intelligence Models: Large Developers," 2025 Reg. Sess. (Cal. 2025) (amended July 17, 2025), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.

care in verifying the statement's truth, particularly when the false information causes reputational harm to a private figure.

The legal calculus becomes more complicated when the harmful output originates autonomously from the AI system, absent user intent to defame. In such cases, courts must confront the question of whether AI developers or platform providers can or should be held liable for speech generated by systems they created or operate. The issue is doctrinally novel, in part because AI lacks the mental state or fault traditionally required in tort law; in addition, at least in some instances, developers may not reasonably foresee specific outputs from models trained on vast and dynamic datasets and responding to myriads of user prompts, where even subtle differences in wording might generate different outputs.

As AI systems grow increasingly sophisticated and autonomous, courts and policymakers must address whether and under what circumstances AI developers or deployers can be held liable for the content their systems generate. Potential factors that may influence liability include the following: the degree of human involvement in the generation and dissemination of the output; the foreseeability of the harmful content; the degree of control or curation exercised by the developer or platform; whether the developer or platform engaged in negligent design, deployment, or moderation practices; and the extent to which the output is understood by ordinary users as factual, given the known propensity of AI systems to "hallucinate" or generate inaccurate information.[54] In this context, practices such as "red teaming" and reinforcement learning from human feedback (RLHF) may become key indicators of whether developers took reasonable steps to anticipate and mitigate foreseeable harms. Their use or omission could inform assessments of negligence or care in high-risk deployments.

An important consideration in assessing defamation liability for AI-generated content is the widely recognized phenomenon that LLMs frequently "hallucinate," producing fabricated information without intent or factual grounding. Given growing public awareness that AI outputs may be unreliable or speculative, courts are increasingly viewing such statements as less likely to be interpreted by a reasonable person as factual assertions, which is a core element of defamation. This understanding was reflected in *Walters v. OpenAI,* where the Superior Court of Gwinnett County, Georgia, granted summary judgment in favor of OpenAI, underscoring the difficulty of sustaining defamation claims involving AI outputs. The lawsuit was brought by a Georgia radio host alleging that ChatGPT falsely claimed he had embezzled funds from a nonprofit. The court found that "a reasonable reader would not have understood" ChatGPT's statements as factual assertions and that the plaintiff, a public figure, failed to demonstrate "knowing or reckless falsehood." It also held that Walters could not show "even negligence," nor provide evidence of "actual damages," all of which are required elements for a libel claim regarding a matter of public concern.[55]

Another high-profile example is the defamation lawsuit filed by political activist Robby Starbuck against Meta, which was settled after alleging that Meta's AI platform produced false and defamatory statements about him in response to user prompts.[56] As part of the settlement, Starbuck will work with Meta to address "ideological and political bias" in its AI.[57] Similarly, in *Battle v. Microsoft,* the plaintiff claimed Bing's AI

54 Amber Bolden, "From Biased Data Inputs to Your Discriminatory Diagnosis Outputs: A Review of Legal Liability for Artificial Intelligence in Healthcare," *Michigan Technology Law Review* 30, no. 2 (2024), https://repository.law.umich.edu/mtlr/vol30/iss2/3.

55 Richard Epstein, "Suing OpenAI for ChatGPT-Produced Defamation Is a Futile Endeavor," *American Enterprise Institute,* January 8, 2025, https://www.aei.org/technology-and-innovation/suing-openai-for-chatgpt-produced-defamation-a-futile-endeavor/; Eugene Volokh, "OpenAI Wins Libel Lawsuit Brought by Gun Rights Activist Over Hallucinated Embezzlement Claims," Reason, May 20, 2025, https://reason.com/volokh/2025/05/20/openai-wins-libel-lawsuit-brought-by-gun-rights-activist-over-hallucinated-embezzlement claims.

56 Sarah Nassauer and Jacob Gershman, "Activist Robby Starbuck Sues Meta Over AI Answers About Him," *Wall Street Journal,* April 29, 2025, https://www.wsj.com/tech/ai/activist-robby-starbuck-sues-meta-over-ai-answers-about-him-9eba5d8a.

57 Joseph De Avila, "Meta, Robby Starbuck Settle AI Defamation Lawsuit," *Wall Street Journal,* August 8, 2025, https://www.wsj.com/tech/ai/meta-robby-starbuck-ai-lawsuit-settlement-6c6e9b0a

defamed him by falsely associating him with a convicted terrorist, though the case was sent to arbitration.[58] At the time of writing, we are not aware of any US court awarding damages in a defamation case involving AI-generated speech.

In the absence of legislative clarity, these questions remain unsettled. Courts adjudicating defamation claims involving AI-generated speech will be tasked with navigating a legal regime that was not designed for autonomous content generation, while balancing the rights of speakers, developers, and injured parties under the constraints of constitutional doctrine.

2.3.2. Section 230 and Platform Immunity

Further complicating the liability landscape is Section 230 of the Communications Decency Act, which provides broad immunity to online platforms for content generated by third parties.[59] This provision has long shielded internet platforms from defamation claims arising from user-generated content.

Whether this protection extends to AI-generated outputs is now the subject of significant legal debate. Courts have begun to consider whether platforms deploying generative AI tools qualify as the "information content providers" of the resulting content, which would open the door to these platforms being held liable.[60] Courts have recognized a limit: They may treat a platform as an information content provider if it "materially contributes" to the development of unlawful content.[61] Under the material contribution test, a provider loses immunity if it is responsible — in part or in whole — for shaping the content's illegality.[62] Thus, if a platform is found to have "materially contributed" to the development of defamatory speech through algorithmic design, prompt structuring, or model fine-tuning, it may lose protections afforded by Section 230. The authors of Section 230 have explicitly stated that AI chatbots would not be shielded by this provision.[63]

This means the applicability of Section 230 in a lawsuit challenging a specific AI-generated output would likely depend on the particular legal claim and the relevant facts. As one group of scholars suggests, generative AI products can be seen as existing on a spectrum, ranging from a retrieval search engine (which is more likely to be covered by Section 230) to a creative engine (which is less likely to be covered).[64] Consequently, Section 230's applicability could differ based on the type of generative AI product, its use cases, and the specific legal claims made.[65]

AI-generated content reflects a form of editorial discretion, shaped by model fine-tuning, red teaming, feedback mechanisms, policy guidelines, and prompt engineering. This type of discretion has long been protected under the First Amendment and is foundational to a functioning digital ecosystem. As generative AI extends the ecosystem beyond traditional platforms like social media and search engines, the absence

58 Battle v. Microsoft Corporation, No. 1:23-cv-01822-LKG (D. Md. Oct. 23, 2024), Memorandum Opinion, https://law.justia.com/cases/federal/district-courts/maryland/mddce/1:2023cv01822/540279/48.

59 Congress.gov, "Section 230 Immunity and Generative Artificial Intelligence," December 28, 2023, https://www.congress.gov/crs-product/LSB11097. Specifically, Section 230(c)(1) states that "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." 47 U.S. Code § 230, https://www.law.cornell.edu/uscode/text/47/230.

60 Noor Waheed, "Section 230 and Its Applicability to Generative AI: A Legal Analysis," *Center for Democracy & Technology*, September 4, 2024, https://cdt.org/insights/section-230-and-its-applicability-to-generative-ai-a-legal-analysis.

61 Fair Housing Council v. Roommates.com, LLC, 521 F.3d 1157, 1166, 1173-74 (9th Cir. 2008); FTC v. Accusearch, Inc., 570 F.3d 1187, 1200 (10th Cir. 2008).

62 Congress.gov, "Section 230: A Brief Overview," August 28, 2025, https://www.congress.gov/crs-product/IF12584.

63 Cristiano Lima-Strong, "AI Chatbots Won't Enjoy Tech's Legal Shield, Section 230 Authors Say," *Washington Post*, March 17, 2023, https://www.washingtonpost.com/politics/2023/03/17/ai-chatbots-wont-enjoy-techs-legal-shield-section-230-authors-say.

64 Peter Henderson, Tatsunori Hashimoto, and Mark A. Lemley, "Where's the Liability in Harmful AI Speech?," *Journal of Free Speech Law* 3, no. 1 (2023): 589-650, https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf#page=1.

65 Congress.gov, "Section 230 Immunity and Generative Artificial Intelligence."

of Section 230 protections removes the statutory shield that has historically enabled diversity and spurred innovation in design choices and content moderation.

2.4. Explicit Content

2.4.1. AI-Generated Child Sexual Abuse Material

2.4.1.1. Federal Laws

A 2023 investigation by Stanford's Internet Observatory identified known child sexual abuse material (CSAM) within a popular open-source dataset, LAION-5B, used to train powerful image-generation models, including Midjourney and Stable Diffusion 1.5.[66] In response to the findings, LAION temporarily took down the dataset to ensure compliance with safety standards.[67] The fact that widely deployed models were trained on such tainted data raised serious concerns about the potential for these tools to inadvertently reproduce illegal content.

There is strong legal consensus in the United States that CSAM involving real minors is not protected by the First Amendment, irrespective of how it is created. The legal status of AI-generated or computer-edited CSAM that does not depict actual children is more complicated. The Supreme Court held that purely virtual or synthetic depictions of children are protected speech unless they are legally obscene under the so-called Miller standard.[68] This standard considers whether "the average person, applying contemporary adult community standards, finds that the matter, taken as a whole, appeals to prurient interests"; "[w]hether the average person, applying contemporary adult community standards, finds that the matter depicts or describes sexual conduct in a patently offensive way"; and "[w]hether a reasonable person finds that the matter, taken as a whole, lacks serious literary, artistic, political, or scientific value."[69]

Under US federal law, computer-generated CSAM may be criminalized if it is indistinguishable from that of a real minor engaged in sexually explicit conduct.[70] Moreover, any visual depiction that is, or appears to be, of a minor engaged in sexually explicit conduct and is obscene can be prosecuted.[71] However, if no real child was involved, if the image is clearly fictional or stylized, and if it fails to meet the Miller obscenity standard, it is generally protected under the First Amendment.

The TAKE IT DOWN Act, passed nearly unanimously by Congress and signed into law by President Trump in May 2025, prohibits the distribution of AI-generated CSAM.[72] The TAKE IT DOWN Act includes nude images published with the intent to "abuse, humiliate, harass, or degrade" a minor rather than only "sexually explicit" images.[73] The law mandates that large online platforms establish a process for victims to report such distribution and strengthens notice-and-reporting mechanisms, which in turn increases the risk that AI companies could be found liable if they knowingly or negligently allow their tools to be used for CSAM creation or distribution. The federal law does not explicitly prohibit personal possession, and U.S. District Judge James

66 David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse," Cyber Policy Center, Stanford University, December 20, 2023, https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.
67 "Safety Review for LAION 5B," LAION.ai, December 19, 2023, https://laion.ai/notes/laion-maintenance.
68 Ashcroft v. Free Speech Coalition, 535 U.S. 234 (2002).
69 U.S. Department of Justice, Criminal Division, "Citizen's Guide to U.S. Federal Law on Obscenity," accessed August 13, 2025, https://www.justice.gov/criminal/criminal-ceos/citizens-guide-us-federal-law-obscenity.
70 18 U.S. Code § 2252A (2018), https://www.law.cornell.edu/uscode/text/18/2252A.
71 18 U.S. Code § 1466A (2018), https://www.law.cornell.edu/uscode/text/18/1466A.
72 Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks (TAKE IT DOWN) Act, Pub. L. No. 119-12, S. 146, 119th Cong., 1st sess. (introduced January 9, 2025, signed into law May 19, 2025), https://www.congress.gov/bill/119th-congress/senate-bill/146/text.
73 Sunny Gandhi and Adam Billen, "The US Senate's Passage of the TAKE IT DOWN ACT Is Progress on an Urgent, Growing Problem," Tech Policy Press, February 21, 2025, https://techpolicy.press/the-us-senates-passage-of-the-take-it-down-act-is-progress-on-an-urgentgrowing-problem.

Peterson ruled that possessing "virtual child pornography" was protected by the First Amendment.[74] While the law addresses an unquestionably serious harm, its expansive enforcement mechanism and vague provisions raise substantial free expression concerns, particularly about how such powers could be used to take down constitutionally protected speech.[75]

2.4.1.2. State Laws

Prior to the passage of the TAKE IT DOWN Act, state legislatures moved swiftly to address CSAM. According to Public Citizen's legislation tracker and research from the advocacy organization Enough Abuse, as of late August 2025, 45 states have enacted laws addressing AI-generated intimate deepfakes that cover minors[76] and criminalizing AI-generated or computer-generated CSAM.[77] These statutes reflect definitive concern about the use of AI to produce exploitative imagery and abuse of children, particularly as such content spreads rapidly across digital platforms.

States such as California and Illinois have enacted robust statutes that unambiguously include computer-generated content within the definition of CSAM. Montana's HB 82 criminalizes the production, distribution, and possession of computer-generated CSAM, regardless of whether a real child was involved in the content's creation.[78] Some states — such as Colorado — use broader language prohibiting "digitally reproduced" visual material, which may not be interpreted to include AI-synthesized content unless judicially construed or legislatively clarified. Nebraska, by contrast, explicitly prohibits "digital image or computer displayed image … whether made or produced by electronic, mechanical, computer or digital or other means," demonstrating more definitive statutory language.[79] Several states criminalize CSAM materials only if they depict a real, identifiable child, while others — such as Texas and Utah — extend criminal liability to any image that reasonably appears to depict a minor engaged in sexual conduct.

2.4.2. AI-Generated Non-Consensual Intimate Imagery

2.4.2.1. Federal Laws

At the federal level, deepfake pornography and AI-generated intimate forgeries have been the subject of increased legislative activity. The TAKE IT DOWN Act requires platforms to take down AI-generated non-consensual intimate imagery (NCII) within 48 hours upon request.[80] While the law responds to emerging forms of digital exploitation, it raises important questions about intermediary liability, platform duties, and the permissible scope of content moderation. As courts have previously cautioned, laws targeting harmful but expressive content must be narrowly tailored and sufficiently clear to avoid restrictive chilling effects on protected speech.

74 Ben Goggin, "Possession of AI-Generated Child Sexual Abuse Imagery May Be Protected by First Amendment in Some Cases, Judge Rules," *NBC News*, March 18, 2025, https://www.nbcnews.com/tech/tech-news/ai-generated-child-sexual-abuse-imagery-judge- ruling-rcna196710.
75 "State Laws Criminalizing AI-Generated or Computer-Edited CSAM," Enough Abuse, n.d., accessed September 5, 2025, https://enoughabuse.org/get-vocal/laws-by-state/state-laws-criminalizing-ai-generated-or-computer-edited-child-sexual-abuse-material-csam.
76 "Tracker: State Legislation on Intimate Deepfakes," Public Citizen, accessed September 5, 2025, https://www.citizen.org/article/tracker-intimate-deepfakes-state-legislation.
77 "State Laws Criminalizing AI-Generated or Computer-Edited CSAM."
78 H.B. 82, "An Act Generally Revising Crimes Against Children; Creating the Offense of Grooming of a Child for a Sexual Offense," 69th Leg (Mont. 2025) (signed by Governor Apr. 7, 2025; effective July 1, 2025), https://bills.legmt.gov/#/laws/bill/2/LC0232?open_tab=bill.
79 "State Laws Criminalizing AI-Generated or Computer-Edited CSAM."
80 TAKE IT DOWN Act, Pub. L. No. 119-12 (2025).

The TAKE IT DOWN Act addresses genuinely serious harms, particularly those facing women, minors, and LGBTQ+ individuals; however, civil liberties groups have raised concerns about its breadth. Future of Free Speech experts have pointed out that the act responds to real harms, but in the hands of a government increasingly willing to regulate speech, its broad provisions provide a powerful tool for censoring lawful expression, monitoring private communications, and undermining due process.[81] The Center for Democracy and Technology (CDT) has warned that without narrowly tailored exemptions the bill could inadvertently criminalize constitutionally protected speech, including artistic, educational, or political content deemed "obscene" or "indecent" by subjective standards.[82] In his March address to a joint session of Congress, President Trump stated, "I'm going to use that bill for myself too, if you don't mind, because nobody gets treated worse than I do online, nobody."[83] President Trump's public endorsement of the bill, coupled with his statement suggesting it could be used to silence critics, has heightened fears of viewpoint-based enforcement and chilling effects.

2.4.2.2. State Laws

Prior to the TAKE IT DOWN Act, states passed a flurry of legislation to address NCII. According to Public Citizen's legislation tracker, as of late August 2025, 41 states have enacted laws addressing AI-generated intimate deepfakes, either by amending existing NCII or "revenge porn" laws or by enacting stand-alone statutes.[84] The TAKE IT DOWN Act has provided a federal net criminalizing both authentic and computer-generated NCII, piecing together the fragmented legal landscape of inconsistent protections and enforcement across state jurisdictions.

Jurisdictions such as California, New York, Virginia, Texas, and Minnesota provide civil and/or criminal remedies for the unauthorized distribution of synthetic sexually explicit images; however, the key provisions vary across state laws. New York and California have both civil remedies and criminal penalties for knowingly distributing deepfake pornography. Utah amended its Sexual Exploitation Act to define "counterfeit intimate image" in a way that expressly includes AI-generated representations, and Indiana has criminalized the distribution of intimate images, AI-generated or otherwise, without the subject's consent.[85]

Several states classify the nonconsensual sharing of deepfake nudes as a form of harassment. In 2024, Massachusetts passed An Act to Prevent Abuse and Exploitation, criminalizing not only traditional "revenge porn" but also the distribution of "digitized" sexually explicit content that appears realistic to a reasonable viewer.[86] Colorado has created a cause of action for nonconsensual disclosure of an intimate digital depiction or threatening to disclose a highly realistic but false visual depiction that has been created, altered, or produced by generative AI or similar tools.[87]

81 Ashkhen Kazaryan and Ashley Haek, "The Road to Enforcement Chaos: The Hidden Dangers of the TAKE IT DOWN Act," *The Bedrock Principle*, May 12, 2025, https://www.bedrockprinciple.com/p/the-road-to-enforcement-chaos-the.
82 Center for Democracy and Technology et al., "Letter Expressing Concerns Regarding the TAKE IT DOWN Act," CDT, February 12, 2025, https://cdt.org/wp-content/uploads/2025/02/TAKE-IT-DOWN-Sign-On-Letter_21225.pdf.
83 Donald J. Trump, "Presidential Address to a Joint Session of Congress," March 4, 2025, C-SPAN, video, https://www.c-span.org/program/joint-session-of-congress/president-trump-addresses-joint-session-of-congress/656056; "Full Transcript of President Trump's Speech to Congress," *New York Times*, March 4, 2025, https://www.nytimes.com/2025/03/04/us/politics/transcript-trump-speech-congress.html.
84 "Tracker: State Legislation on Intimate Deepfakes," Public Citizen, accessed September 5, 2025, https://www.citizen.org/article/tracker-intimate-deepfakes-state-legislation.
85 "State Laws Criminalizing AI-Generated or Computer-Edited CSAM."
86 "State Laws Criminalizing AI-Generated or Computer-Edited CSAM.
87 S.B. 288, "Intimate Digital Depictions Criminal & Civil Actions," 75th Gen. Assemb. (Colo. 2025) (signed by Governor June 2, 2025), https://leg.colorado.gov/bills/sb25-288.

2.5. Hate Speech

The First Amendment provides some of the most robust protections for freedom of expression in the world, extending even to speech that is grossly offensive or hateful. Unlike many democracies that criminalize certain forms of hate speech, the United States has no general statutory prohibition on hate speech. The Supreme Court has consistently rejected government efforts to restrict speech based solely on its hateful or offensive nature. The court has held that even inflammatory speech is protected unless it is intended and likely to incite imminent lawless action, which is a high bar that continues to limit government regulation.[88] The court has emphasized that "[t]he government may not regulate speech based on hostility — or favoritism — towards the underlying message expressed."[89] The First Amendment does not contain a hate speech exception, and courts have reaffirmed that offensive expression is not a sufficient basis for state censorship.[90]

As applied to AI-generated hate speech, this constitutional principle presents significant constraints on government regulation. AI-generated expression, even when offensive or derogatory, would likely be protected unless it falls into one of the narrow, historically recognized categories of unprotected speech, such as incitement to imminent lawless action,[91] true threats,[92] or obscenity.[93] Accordingly, broad governmental attempts to regulate or ban AI-generated hate speech face serious constitutional challenges, particularly if based on the viewpoint or content of the speech itself.

The private sector is not bound by the First Amendment, allowing AI developers and platform operators to design and enforce their own content moderation policies — such as acceptable use policies or fine-tuning practices — that filter out hate speech or other forms of offensive content. Many platforms employ these measures as part of corporate social responsibility initiatives or to comply with global norms and user expectations.

Reliance on automated moderation systems for detecting hate speech raises inherent difficulties, as definitions differ over which groups are "protected," how severity is assessed, and the potential for restricting speech that is merely offensive, satirical, or part of legitimate discussion. These ambiguities create a significant risk of over-removal — where lawful, socially valuable expression is inadvertently suppressed. For example, a user might ask a chatbot to summarize historical writings or political rhetoric that contains offensive language; while the material may be unpleasant, it could serve an educational or research purpose in context. Where chatbot interactions are private, there is a strong case for allowing more speech than on public platforms, such as social media. Overly broad filters in LLMs can chill inquiry, suppress satire, and erase legitimate political commentary. As demonstrated in our previous report, *A Snapshot of Content Policies*, opaque automated moderation and overinclusive policies can magnify these harms, underscoring the need for narrowly defined rules for restricting expression.[94]

Recent state-level efforts to mandate transparency in platform content moderation, particularly around hate speech and disinformation, underscore the legal tension between regulating harmful content and preserving First Amendment rights. For example, laws in California and New York have sought to compel platforms

88 Brandenburg v. Ohio, 395 U.S. 444 (1969).
89 R.A.V. v. City of St. Paul, 505 U.S. 377 (1992).
90 Snyder v. Phelps, 562 U.S. 443 (2011).
91 Brandenburg v. Ohio, 395 U.S. 444 (1969).
92 Counterman v. Colorado, 600 U.S. 66 (2023).
93 Miller v. California, 413 U.S. 15 (1973).
94 Jordi Calvet-Bademunt and Jacob Mchangama, "Freedom of Expression in Generative AI: A Snapshot of Content Policies," The Future of Free Speech, February 2024, https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.

to disclose their definitions and policies for moderating hate speech.[95] However, courts have found that such mandates may infringe on editorial discretion and amount to compelled speech. These legal setbacks highlight the constitutional limits on government attempts to influence how private actors address hate speech online, even indirectly.

2.6. Election and Political Content

2.6.1. Constitutional Protection of AI-Generated Deepfakes

The US Supreme Court has long held that political speech lies at the heart of First Amendment protections, even when such speech is demonstrably false. In *United States v. Alvarez* (2012), the court struck down the Stolen Valor Act, reaffirming that the government cannot categorically prohibit false speech unless it causes a legally cognizable harm or falls within a historically unprotected category.[96] As such, false political speech, including AI-generated disinformation, retains robust constitutional protection unless it amounts to defamation, incitement, or fraud.[97] This broad constitutional shield limits public authorities' ability to regulate AI-generated political content, especially where such efforts resemble prior restraints or content-based restrictions. Sweeping restrictions on deepfakes without clear, narrow definitions and safeguards may chill lawful expression, discourage public-interest uses of synthetic media, and deter innovation in political communication technologies.

In early 2024, an AI-generated robocall impersonating President Biden urged New Hampshire voters to skip the state's primary election.[98] In response, the Federal Communications Commission issued a declaratory ruling clarifying that AI-generated voice clones in robocalls qualify as "artificial" under the Telephone Consumer Protection Act, thereby subjecting them to federal restrictions.[99] This regulatory move focused on the method of communication rather than the content of the message, highlighting the limited avenues available to address deceptive political speech without triggering First Amendment concerns.

Attempts to ban or suppress political deepfakes — defined as digitally altered media impersonating real individuals in campaign contexts — have occurred at the state level. They are often met with First Amendment challenges, as in the case involving California's AB 2839,[100] which sought to restrict AI-generated deepfakes during elections and was struck down on First Amendment grounds. US District Judge John Mendez stated "AB 2839 suffers from 'a compendium of traditional First Amendment infirmities,' stifling too much speech while at the same time compelling it on a selective basis … When it comes to political expression, the antidote is not prematurely stifling content creation and singling out specific speakers but encouraging counter speech, rigorous fact-checking, and the uninhibited flow of democratic discourse. California cannot pre-emptively sterilize political content."[101]

95 A.B. 587, "Social Media Companies: Terms of Service," 2021–2022 Reg. Sess. (Cal. 2022) (approved by Governor Sept. 13, 2022), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB587; Nick Robins-Early, "Elon Musk's X Sues New York over Hate Speech and Disinformation Law," *The Guardian*, June 17, 2025, https://www.theguardian.com/technology/2025/jun/17/elon- musk-new-york-hate-lawsuit-speech-law.

96 United States v. Alvarez, 567 U.S. 709 (2012).

97 Rod Kubat, "Constitutional Free Speech Protection of Lies in Political Campaigns," American Bar Association, September 2024, https://www.americanbar.org/groups/senior_lawyers/resources/voice-of-experience/2024-september/constitutional-free-speech-protection-of-lies-in-political-campaigns.

98 Holly Ramer and Ali Swenson, "Political Consultant Behind Fake Biden Robocalls Faces $6 Million Fine and Criminal Charges," AP News, May 23, 2024, https://apnews.com/article/biden-robocalls-ai-new-hampshire-charges-fines-9e9cc63a71eb9c78b9bb0d1ec2aa6e9c.

99 Federal Communications Commission, "FCC Makes AI-Generated Voices in Robocalls Illegal," Declaratory Ruling, FCC-24-17A1, Feb. 8, 2024, https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal.

100 A.B. 2839, "Elections: Deceptive Media in Advertisements," 2023–24 Reg. Sess. (Cal. 2024) (approved by Governor Sept. 17, 2024; chap. 262), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2839.

101 Kohls v. Bonta, Case No. 2:24-cv-02527-JAM-CKD, Order Granting Plaintiff's Motion for Summary Judgment as to AB 2839 (E.D. Cal. Aug. 29, 2025), 23; Washington Post v. McManus, 944 F.3d 506 (4th Cir. 2019).

Another California law, Defending Democracy from Deepfake Deception Act of 2024 (AB 2655), required disclosures on social media and empowers platforms to label or block synthetic media used in a political context.[102] However, Judge Mendez struck down this law on Section 230 grounds and declined to address the free speech arguments presented.[103] Given California's leadership in AI regulation, these rulings may provide a shield against similar legislative efforts in other states, especially where courts are already scrutinizing such laws on constitutional grounds.

This US approach stands in sharp contrast to those in other countries, where publishing false information can lead to harsh punishments and where the legal threshold for restricting such speech is far lower.[104] For example, Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA) enables authorities to tackle fake news and can result in fines and imprisonment of up to five years, with penalties doubled if the individual used bots for spreading what the government deems false statements against the public interest.[105] South Korea enacted amendments in 2024 that criminalize all election-related deepfakes during the 90-day period before elections, with violations punishable by imprisonment of up to seven years or by a fine of up to 50 million won.[106] These countries may frame deepfakes as existential threats to electoral integrity, justifying sweeping controls that are constitutionally unthinkable in the United States. In the United States, the "elite panic" over deepfakes and elections has largely failed to materialize. Despite high-profile incidents like the Biden robocall, there is little evidence that synthetic media has meaningfully altered electoral outcomes.[107] By preserving robust First Amendment protections, the United States avoids reflexive overregulation and ensures that the tools used to address genuine harms do not become blunt instruments for suppressing political dissent, satire, or inconvenient truths.

2.6.2. State-Level Legislative Efforts on Deepfakes

Despite these constitutional hurdles, at least 28 states have enacted laws regulating AI-generated political deepfakes, with another 13 states considering similar measures as of late August 2025.[108] These statutes adopt one of two approaches: mandatory disclosure requirements or temporal prohibitions on deceptive content.

2.6.2.1. Political Communication Disclosures

Several states — California, Michigan, Utah, Alabama, Arizona, and Oregon — have adopted laws requiring clear and conspicuous disclosures on political advertisements or communications that involve synthetic or manipulated media. These laws often impose such requirements within a specific window preceding an election and may include formatting standards for disclaimers or mandates for metadata tagging.

For instance, Michigan mandates disclosures for AI-modified ads, while Utah requires labeling for synthetic content and metadata obligations. In April 2025, North Dakota introduced new regulations for the use of AI in

102 A.B. 2655, "Defending Democracy from Deepfake Deception Act of 2024," 2023-24 Reg. Sess. (Cal. 2024) (approved by Governor Sept. 17, 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2655.
103 Kohls v. Bonta, Case No. 2:24-cv-02527-JAM-CKD, Order and Final Judgment and Permanent Injunction as to AB 2655 (E.D. Cal. Aug. 20, 2025); Chase DiFeliciantonio, "Elon Musk and X Notch Court Win Against California Deepfake Law," *Politico*, August 5, 2025, https://www.politico.com/news/2025/08/05/elon-musk-x-court-win-california-deepfake-law-00494936.
104 Gabrielle Lim and Samantha Bradshaw, "Chilling Legislation: Tracking the Impact of 'Fake News' Laws on Press Freedom Internationally," National Endowment for Democracy, July 19, 2023, https://www.cima.ned.org/publication/chilling-legislation.
105 "Singapore: 'Fake News' Law Curtails Speech," *Human Rights Watch*, January 13, 2021, https://www.hrw.org/news/2021/01/13/singapore-fake-news-law-curtails-speech.
106 Tae Yeon Eom, "South Korea Contends with AI and Electoral Integrity," East Asia Forum, April 1, 2024, https://eastasiaforum.org/2024/04/01/south-korea-contends-with-ai-and-electoral-integrity.
107 Sayash Kapoor and Arvind Narayanan, "We Looked at 78 Election Deepfakes: Political Misinformation Is Not an AI Problem," Knight First Amendment Institute at Columbia University, December 13, 2024, https://knightcolumbia.org/blog/we-looked-at-78-electiondeepfakes-political-misinformation-is-not-an-ai-problem; Sam Stockwell et al., "AI-Enabled Influence Operations: Safeguarding Future Elections," Centre for Emerging Technology and Security, Alan Turing Institute, November 13, 2024, https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections.
108 "Tracker: State Legislation on Deepfakes in Elections," Public Citizen, accessed September 5, 2025, https://www.citizen.org/article/ tracker-legislation-on-deepfakes-in-elections.

political communications, specifically that "any political content that uses AI to visually or audibly impersonate a human must prominently display [a] disclaimer."[109]

Although disclaimer requirements are viewed as less restrictive than outright bans, they remain subject to First Amendment scrutiny. Courts have upheld similar disclosure mandates in the campaign finance context,[110] but concerns about compelled speech persist.[111] Forced disclosure laws can infringe on speakers' autonomy by compelling them to include disclaimers they may not agree with, altering their intended message. Such mandates may also chill protected expression, as speakers might avoid using AI-generated content altogether to sidestep compliance burdens, legal risks, or public skepticism. This deterrent effect is especially concerning in political and artistic contexts, where vague or overbroad definitions of "synthetic" content can lead to self-censorship. Required disclaimers may stigmatize the underlying message, signaling to audiences that it is less credible or inherently misleading, even when the content is lawful and constitutionally protected.

2.6.2.2. Political Deepfake Prohibitions

Some states have adopted outright prohibitions on the dissemination of political deepfakes, particularly close to elections. Minnesota and Texas criminalize the publication of materially deceptive political media within a defined pre-election window.[112] South Dakota prohibits undisclosed deepfakes within 90 days of an election, subject to an affirmative defense if proper disclosures are made.[113] Kentucky's legislation permits remedies for candidates harmed by synthetic media and includes additional provisions regulating "high-risk AI systems" used in political decision-making.[114]

Legal challenges to these statutes often center on overbreadth and vagueness, as well as failure to distinguish harmful manipulation from protected satire and parody. Minnesota's statute prohibits deepfakes intended to "injure" a candidate or "influence" an election.[115] This law is currently being challenged in federal court on similar grounds as California's (recently struck down) deepfake laws,[116] including vagueness and potential conflicts with Section 230. The plaintiff, X, argues that the law's requirements are so unclear that social media platforms cannot understand what is permitted or prohibited, potentially leading to over-censorship of valuable political speech. As this example shows, even well-intentioned statutes can backfire from being too imprecise and susceptible to abuse — curbing public debate, suppressing diverse political viewpoints, and undermining the very democratic values they aim to protect. Regulation should not sacrifice the open exchange of ideas that is essential to a functioning democracy, and any restrictions should be limited and address only real, direct, and imminent harms, which does not yet include political deepfakes.

These types of outright bans raise significant First Amendment concerns because they restrict speech based on content, timing, and intent — each of which triggers heightened constitutional scrutiny.[117] Laws that

109 H.B. 1167, "An Act to Create and Enact a New Section to Chapter 16.1-10 of the North Dakota Century Code, Relating to Artificial Intelligence Disclosure Statements," 69th Leg. Assemb. (N.D. 2025) (signed by Governor Apr. 11, 2025; filed with Secretary of State Apr. 11, 2025), https://ndlegis.gov/assembly/69-2025/regular/bill-overview/bo1167.html.
110 Citizens United v. FEC, 558 U.S. 310 (2010).
111 R. Sam Garrett, "The State of Campaign Finance Policy: Recent Developments and Issues for Congress," Congressional Research Service Report R41542, July 29, 2025, https://www.congress.gov/crs-product/R41542; Wooley v. Maynard, 430 U.S. 705 (1977).
112 Chris McIsaac, "Update on 2025 State Legislation to Regulate Election Deepfakes," R Street Institute, March 17, 2025, https://www.rstreet.org/commentary/update-on-2025-state-legislation-to-regulate-election-deepfakes.
113 S.B. 164, "An Act to Prohibit the Use of a Deepfake to Influence an Election and to Provide a Penalty Therefor," 2025 Leg. (S.D. 2025) (signed by Governor Mar. 25, 2025; S.J. 539), https://legiscan.com/SD/text/SB164/id/3165619.
114 S.B. 4, "An Act Relating to Protection of Information and Declaring an Emergency," 2025 Leg. (Ky. 2025) (signed by Governor Mar. 24, 2025; Acts Ch. 66), https://apps.legislature.ky.gov/record/25RS/sb4.html.
115 H.F. 4772, "Omnibus Elections Policy Bill," 93rd Leg., Reg. Sess. (Minn. 2024) (signed by Governor May 17, 2024; filed with Secretary of State May 20, 2024), https://www.revisor.mn.gov/laws/2024/0/112/laws.2.76.0#laws.2.76.0.
116 Steve Karnowski, "Elon Musk's X Sues to Overturn Minnesota Political Deepfakes Ban," *AP News*, April 25, 2025, https://apnews.com/article/minnesota-deepfake-law-x-elon-musk-twitter-c423540850ca3837891d62d69c6639f1.
117 Reed v. Town of Gilbert, 576 U.S. 155 (2015); Citizens United v. FEC, 558 U.S. 310 (2010); and FCC v. Fox Television Stations, 567 U.S.239 (2012).

criminalize the dissemination of "materially deceptive" or "injurious" content without clear definitions risk sweeping under their purview legitimate political critique, parody, or satire, which are common features of campaign discourse. The lack of clear standards may also cause platforms and speakers to over-censor to avoid liability, chilling lawful expression. As a result, even well-intentioned efforts to combat misinformation can backfire by curbing public debate and suppressing diverse political viewpoints at critical moments in the democratic process.

2.6.2.3. Definitions and Enforcement Mechanisms

One obstacle to uniform regulation is the lack of consensus on definitions. State laws variably refer to "deepfakes," "synthetic media," and "deceptive media," with differing thresholds for intent, scope, and technology covered. Some focus exclusively on video content, while others include audio- and text-based manipulations.[118] Enforcement mechanisms also vary, with laws providing civil injunctive relief, statutory damages, or criminal penalties. Although most of these laws have exemptions for satire, parody, and journalism, these exemptions may not fully insulate protected speech in practice.

2.7. Copyright

2.7.1. Use of Copyrighted Material in AI Training

The use of copyrighted content as training data for AI models has emerged as a defining legal question in the governance of generative technologies. Central to this dispute is whether the ingestion of copyrighted works by AI systems, particularly LLMs, without a license constitutes infringement or falls within the bounds of the fair use doctrine.[119] Proponents of permissibility argue that training constitutes a transformative use because it does not reproduce the original expression but instead contributes to the creation of new outputs that are not copies of the input data.[120] This argument is often grounded in the view that training data merely informs a statistical model and does not result in direct substitution or market harm.

Recent litigation has challenged this theory. In *Thomson Reuters v. Ross Intelligence,*[121] the District Court for the District of Delaware rejected a fair use defense in a case involving the use of copyrighted legal headnotes to train a non-generative legal research tool.[122] Although the system at issue was not generative, the decision signals judicial skepticism toward the unlicensed appropriation of copyrighted materials in AI development, particularly where the use is commercial in nature and the input data is reproduced in a non-trivial way.

Courts have begun to diverge in their treatment of fair use claims in the generative AI context. In *Authors Guild v. Anthropic,* US District Judge William Alsup ruled that using copyrighted books to train Anthropic's Claude model qualified as fair use, emphasizing that the use was "quintessentially transformative" because it enabled

118 CJ Larkin, "Regulating Election Deepfakes: A Comparison of State Laws," Tech Policy Press, January 8, 2025, https://techpolicy.press/regulating-election-deepfakes-a-comparison-of-state-laws.
119 17 U.S. Code § 107, https://www.law.cornell.edu/uscode/text/17/107.
120 Virginie Berger, "The AI Copyright Battle: Why OpenAI and Google Are Pushing for Fair Use," *Forbes,* March 15, 2025, https://www.forbes.com/sites/virginieberger/2025/03/15/the-ai-copyright-battle-why-openai-and-google-are-pushing-for-fair-use
121 Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., No. 1:20-cv-00613-SB (D. Del. Feb. 11, 2025), https://www.ded.uscourts.gov/sites/ded/files/opinions/20-613_5.pdf.
122 "Court Shuts Down AI Fair Use Argument in Thomson Reuters Enterprise Centre GMBH v. Ross Intelligence Inc.," Reed Smith, March 3, 2025, https://www.reedsmith.com/en/perspectives/2025/03/court-ai-fair-use-thomson-reuters-enterprise-gmbh-ross-intelligence.

the generation of new text rather than reproducing the original works.[123] Still, Judge Alsup allowed the case to proceed on narrower grounds, finding that Anthropic could be liable for storing over seven million pirated books in a centralized library, and ordered a trial to determine damages related to that retention.[124]

By contrast, a lawsuit against Meta was dismissed by US District Judge Vince Chhabria, who found that the plaintiffs — thirteen authors alleging unauthorized use of their books to train Meta's Llama model — had failed to articulate a viable legal theory or present sufficient factual evidence.[125] Notably, Judge Chhabria made clear that his ruling did not determine whether Meta's conduct was lawful, suggesting that more carefully crafted claims could still succeed. These rulings underscore the unsettled nature of fair use jurisprudence in AI and foreshadow continued legal uncertainty over how courts will address the tension between transformative machine learning practices and traditional copyright protections.

The US Copyright Office has released the pre-publication version of Part 3 of its "Copyright and Artificial Intelligence" report, focusing on generative AI training and the applicability of the fair use doctrine.[126] The report details several stages in the development and deployment of general AI models where the use of copyrighted materials for training could implicate copyright protections. The Copyright Office states, "In the Office's view, training a generative AI foundation model on a large and diverse dataset will often be transformative," while noting that this is not absolute.[127] It points out that "making commercial use of vast troves of copyrighted works to produce expressive content that competes with them in existing markets, especially where this is accomplished through illegal access, goes beyond established fair use boundaries."[128] The head of the Copyright Office was fired shortly after releasing the report.[129] At the time of writing, the Copyright Office "recommends allowing the licensing market to continue to develop without government intervention."[130]

2.7.2. Copyrightability of AI-Generated Works

Another fundamental issue in AI law involves the copyright eligibility of works generated wholly or partly by artificial intelligence. US copyright law, as articulated by the Constitution and the Copyright Act of 1976, requires human authorship for a work to be eligible for protection.[131] This principle was affirmed in *Thaler v. Perlmutter,* where the US District Court for the District of Columbia upheld the Copyright Office's refusal to register a visual artwork generated solely by an AI system.[132] The court emphasized that human authorship is a "bedrock requirement" of copyright law.

The Copyright Office reaffirmed this position in Part 2 of "Copyright and Artificial Intelligence," concluding that existing statutory and doctrinal frameworks are sufficient to resolve most issues related to AI-generated outputs.[133] The report draws a bright-line distinction between AI as a creative assistant and AI as the originator

123 Andrew Jeong, "Federal Court Says Copyrighted Books Are Fair Use for AI Training," *Washington Post*, June 25, 2025, https://www.washingtonpost.com/technology/2025/06/25/ai-copyright-anthropic-books

124 Most recently, a US judge has certified "Napster-style" copyright class action against Anthropic. See Emma Whitford, "US Judge Certifies 'Napster-Style' Copyright Class Action Against Anthropic," MLex, July 17, 2025, https://www.mlex.com/mlex/artificial-intelligence/articles/2366395/us-judge-certifies-napster-style-copyright-class-action-against-anthropic.

125 Dan Milmo, "Meta Wins AI Copyright Lawsuit as US Judge Rules Against Authors," *The Guardian*, June 26, 2025, https://www.theguardian.com/technology/2025/jun/26/meta-wins-ai-copyright-lawsuit-as-us-judge-rules-against-authors.

126 US Copyright Office, "Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version)," in Report on *Copyright and Artificial Intelligence* (Washington, DC: US Copyright Office, May 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.

127 US Copyright Office, "Copyright and Artificial Intelligence, Part 3," 45.

128 US Copyright Office, "Copyright and Artificial Intelligence, Part 3," 107.

129 Andrew Limbong, "The U.S. Copyright Office Used to Be Fairly Low-Drama: Not Anymore," *NPR*, June 6, 2025, https://www.npr.org/2025/06/06/nx-s1-5399781/copyright-office-explainer-perlmutter-trump.

130 US Copyright Office, "Copyright and Artificial Intelligence, Part 3," 106.

131 Copyright Law of the United States, Title 17, US Copyright Office (December 2024), https://www.copyright.gov/title17/title17.pdf.

132 Thaler v. Perlmutter, 687 F. Supp. 3d 140, 142 (D.D.C. 2023).

133 US Copyright Office, "Copyright and Artificial Intelligence, Part 2: Copyrightability," in *Report on Copyright and Artificial Intelligence*, (Washington, DC: US Copyright Office, January 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf.

of expression. Where a human meaningfully selects, arranges, or modifies AI-generated elements, the resulting work may qualify for copyright protection. However, content generated autonomously by an AI system without sufficient human authorship is not copyrightable under current law.

Notably, the 2025 report explicitly states that prompts alone, even highly sophisticated ones, do not confer authorship over the AI output. The Copyright Office emphasized that copyright's core purpose of incentivizing and rewarding human creativity does not extend to non-human actors or their outputs, regardless of technological sophistication.

2.7.3. Ongoing Infringement Concerns

Beyond ownership, the potential for infringement through AI-generated outputs poses novel legal questions. Where an AI system produces content that is substantially similar to a copyrighted work in the training dataset, issues of derivative works and unauthorized reproduction arise. Legal scholars emphasize that for an AI-generated output to infringe upon a copyrighted work, it must be "substantially similar" and replicate original elements of the copyrighted work. Courts have generally required that the allegedly infringing work incorporate protected expression from the original work, with mere stylistic resemblance or unprotectable ideas typically falling outside the scope of infringement. Additionally, courts may consider whether the AI-generated content could substitute for the original, potentially harming the market for the original.[134]

Plaintiffs in ongoing litigation have alleged that AI outputs closely mimic the style, structure, or content of protected works, creating risks of substitution and consumer confusion. In the case of *Kadrey v. Meta Platforms, Inc.,* the court dismissed the claim that AI models themselves are infringing derivative works simply because they were trained on copyrighted materials. The court emphasized the necessity to demonstrate that specific outputs incorporate protected elements of the plaintiffs' works.[135] A related issue arose in *Getty Images v. Stability AI,* where Getty alleged that Stability's use of its images infringed its intellectual property rights. Due to jurisdictional challenges, Getty discontinued its primary copyright infringement and database right claims,[136] making the "decision to pursue only the claims for trademark infringement, passing off and secondary infringement of copyright."[137]

Legislative proposals are beginning to respond to these challenges. While the US Copyright Office maintains that the Copyright Act is largely sufficient to address issues of authorship and registration, it has also pointed out that statutory clarification may be needed in adjacent domains, particularly regarding unauthorized digital replicas and the use of AI in impersonation or synthetic likeness generation.[138] Proposals under consideration include measures aimed at regulating the distribution of AI tools designed to reproduce protected content, as well as transparency mandates for developers of generative systems.

---

134 Pamela Samuelson, "Legal Challenges to Generative AI, Part II," Communications of the ACM, November 1, 2023, https://cacm.acm.org/opinion/legal-challenges-to-generative-ai-part-ii; Tori Noble and Mitch Stoltz, "EFF Urges Court to Avoid Fair Use Shortcuts in Kadrey v. Meta Platforms," Electronic Frontier Foundation, April 15, 2025, https://www.eff.org/deeplinks/2025/04/eff-urges-court- avoid-fair-use-shortcuts-kadrey-v-meta-platforms.

135 Kate Knibbs, "A Judge Says Meta's AI Copyright Case Is About 'the Next Taylor Swift,'" *Wired,* May 1, 2025, https://www.wired.com/story/meta-lawsuit-copyright-hearing-artificial-intelligence.

136 As defined by *LexisNexis,* "Primary infringement occurs when a person does, or authorises another to do, any of the restricted acts without the permission of the owner of the copyright … Secondary infringement occurs 'further down the supply chain' where infringing works are dealt with or their production facilitated." "Infringement of Copyright Definition," *LexisNexis,* accessed August 6, 2025, https://www.lexisnexis.co.uk/legal/glossary/infringement-of-copyright. Database right "is an exclusive right which is granted to the maker of a database where there has been a substantial investment in obtaining, verifying or presenting the contents of the database." "Database Right Definition," *LexisNexis,* accessed August 6, 2025, https://www.lexisnexis.co.uk/legal/glossary/database-right-. See also "Passing Off Definition," LexisNexis, accessed August 6, 2025, https://www.lexisnexis.co.uk/legal/glossary/passing-off: "Passing off is a common law action used to protect unregistered trade mark rights in the UK."

137 Kelvin Chan, "Getty Drops Copyright Allegations in UK Lawsuit Against Stability AI," *AP News,* June 25, 2025, https://apnews.com article/getty-images-stability-ai-copyright-trial-stable-diffusion-7208c729fb10c1f133cb49da2065d72a; Sophie Burgess et al., "The UK Getty Trial: Key Takeaways on the AI/Copyright Case," JD Supra, July 9, 2025, https://www.jdsupra.com/legalnews/the-uk-getty-trial- key-takeaways-on-the-5040227.

138 US Copyright Office, "Copyright and Artificial Intelligence, Part 3."

2.7.4. Protections for Digital Likeness and Voice

The federal initiative NO FAKES Act of 2025 would create a federal right protecting an individual's voice and likeness from unauthorized digital replicas.[139] However, in the absence of federal standards, states have legislated in areas tangential to copyright, particularly the unauthorized commercial use of an individual's likeness, voice, or image through generative AI. These laws often draw on the right of publicity doctrine, which protects a person from having their name, image, voice, or other personal features — like a nickname, signature, or photo — used for commercial gain without their permission.

Tennessee enacted the Ensuring Likeness, Voice, and Image Security (ELVIS) Act, which extends protections against the unauthorized use of a person's voice or likeness via synthetic media.[140] Following Tennessee's lead, California,[141] Illinois,[142] Utah,[143] and Arkansas[144] enacted similar laws restricting the nonconsensual use of AI-generated likenesses in commercial or misleading contexts and, in some cases, regulating the tools used to create such replicas.

2.8. Measures Empowering Freedom of Expression

In recognition of the challenges posed by AI-generated misinformation, there is a growing emphasis on increasing public media literacy. Organizations such as the National Association for Media Literacy Education have launched AI literacy initiatives, helping individuals understand how generative AI functions, how it may be used to mislead, and how to verify the credibility of AI content.[145] These educational efforts reflect the constitutional preference for "counterspeech" over censorship, a principle articulated in seminal First Amendment jurisprudence.[146]

President Trump's April 2025 executive order Advancing Artificial Intelligence Education for American Youth directs federal agencies to promote and integrate AI literacy into K–12 curricula and educator training.[147] The initiative established a White House Task Force on AI Education to coordinate efforts and launched a Presidential AI Challenge to encourage and highlight student and educator achievement in AI. Complementing this federal effort, over 60 organizations, including major AI companies like Microsoft, OpenAI, Google, Anthropic, and NVIDIA, have signed a White House "AI Education Pledge," committing to support AI literacy through free tools, curriculum development, grants, and teacher training.[148]

Several private companies have also launched direct initiatives. Microsoft, OpenAI, and Anthropic, in partnership with the American Federation of Teachers, are backing a new National Academy for AI Instruction, which aims to train hundreds of thousands of K–12 educators.[149] The academy will offer workshops and online

139 Nurture Originals, Foster Art, and Keep Entertainment Safe Act of 2025, S. 1367, 119th Cong., 1st sess. (2025,) https://www.congress.gov/bill/119th-congress/senate-bill/1367.
140 H.B. 2091, "An Act to Amend Tennessee Code Annotated, Title 39, Chapter 14, Part 1 and Title 47, Relative to the Protection of Personal Rights," 113th Gen. Assemb. (Tenn. 2024) (signed by Governor Mar. 21, 2024; Pub. Ch. 588, Mar. 26, 2024; effective July 1, 2024), https://www.capitol.tn.gov/Bills/113/Bill/HB2091.pdf.
141 A.B. 2602, "Contracts Against Public Policy: Personal or Professional Services: Digital Replicas," 2023–24 Reg. Sess. (Cal. 2024) (approved by Governor Sept. 17, 2024; chap. 259), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2602; A.B. 1836, "Use of Likeness: Digital Replica," 2023–24 Reg. Sess. (Cal. 2024) (approved by Governor Sept. 17, 2024; chap. 836; effective Jan. 1, 2026), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB1836.
142 Digital Voice and Likeness Protection Act, 815 Ill. Comp. Stat. 550 (2024) (P.A. 103-830, eff. Aug. 9, 2024), https://www.ilga.gov/Legislation/ILCS/Articles?ActID=4531&ChapterID=67.
143 S.B. 271, "Unauthorized Artificial Intelligence Impersonation Amendments," 2025 Gen. Sess. (Utah 2025) (approved by Governor Mar. 27, 2025), https://le.utah.gov/~2025/bills/static/SB0271.html.
144 H.B. 1071, "AI Fairness in Decision-Making Amendments," 2025 Gen. Sess. (Ark. 2025) (signed by Governor Feb. 25, 2025; Act 159), https://arkleg.state.ar.us/Bills/Detail?id=hb1071&ddBienniumSession=2025/2025R.
145 "AI Literacy Initiative," NAMLE, September 9, 2024, https://namle.org/ai-literacy-press-release/.
146 Nadine Strossen, "Counterspeech in Response to Changing Notions of Free Speech," *Human Rights Magazine*, American Bar Association, November 19, 2018, https://www.americanbar.org/groups/crsj/resources/human-rights/archive/counterspeech-response-changing-notions-free-speech.
147 Exec. Order No. 14277, 90 Fed. Reg. 17519 (Apr. 23, 2025), https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth.
148 Ashley Gold, "Exclusive: White House Announces AI Education Pledge," Axios, June 30, 2025, https://www.axios.com/pro/tech-policy/2025/06/30/white-house-announces-ai-education-pledge.
149 Ashley Gold, "Exclusive: White House Announces AI Education Pledge," *Axios*, June 30, 2025, https://www.axios.com/pro/tech-policy/2025/06/30/white-house-announces-ai-

courses to help teachers responsibly integrate AI tools, such as lesson planners and quiz generators, into classrooms, with a focus on transparency, ethics, and privacy. These efforts reflect a growing public-private alignment around making AI education a core part of digital and civic literacy in the United States.

The 2025 AI Action Plan highlights the need to ensure that AI protects free speech — a laudable objective. Free speech advocates should remain vigilant though. The plan frames certain ideological positions — such as DEI initiatives or climate change discourse — as biased, seeks to define neutrality, and emphasizes countering Chinese talking points. By presenting one perspective as the standard of neutrality, the plan risks replacing one orthodoxy with another.[150] The AI Action Plan was accompanied by the executive order Preventing Woke AI in the Federal Government, which, like the plan, invokes the language of free speech while advancing troubling rhetoric and provisions that risk undermining it.[151]

education-pledge.

150 "AFT to Launch National Academy for AI Instruction with Microsoft, OpenAI, Anthropic and United Federation of Teachers," American Federation of Teachers, July 8, 2025, https://www.aft.org/press-release/aft-launch-national-academy-ai-instruction-microsoft-openai-anthropic-and-united.

151 Anzabi and Calvet-Bademunt, "The Anti-'Woke' AI Agenda & Free Speech."

3. Conclusion

The relationship between AI and freedom of expression in the United States is intricate and constantly evolving. While the foundational principles of the First Amendment are likely to extend to AI-generated content, the way they will be applied remains a subject of ongoing debate and legal development. The current federal policy landscape has allowed individual states to address specific alleged harms and concerns arising from AI technologies. States' AI regulations target six core concerns: high-risk AI systems and algorithmic discrimination, disclosure and labeling requirements, frontier model safety, access to computation and accountability, explicit content, and political deepfakes and deceptive media. However, attempts to ban or suppress political deepfakes have led to overly vague and broad restrictions and already face constitutional challenges, highlighting the inherent difficulties in regulating AI-generated speech without infringing on essential First Amendment rights.

The legal status of using copyrighted material for AI training and the copyrightability of AI-generated outputs are also critical areas of contention with ongoing litigation. The issue of liability for AI-generated harmful content — such as defamation, CSAM, and NCII — is being addressed through a combination of existing laws and new legislation. The TAKE IT DOWN Act is a stand-alone example of federal AI regulation. This law addresses an unquestionably serious harm, but its expansive enforcement mechanism and vague provisions raise substantial free expression concerns. Though strong consensus exists regarding sensitive areas, the regulation of other forms of harmful content, such as hate speech and misinformation, will come head-to-head with the robust free speech protections in the United States. Some state laws restricting political deepfakes have already been blocked. Additionally, the nation is seeing a federal and private push toward greater free speech protections in the AI landscape, along with measures to empower continued AI adoption.

Ultimately, safeguarding freedom of expression in the age of AI requires embedding robust, speech-protective principles into law and policy — principles that limit restrictions to addressing only real, direct, and imminent harms. Policymakers, legal scholars, and technology developers should focus on ensuring that AI's transformative capabilities remain a force for expanding, not constraining, free expression. This approach recognizes that AI not only is shaped by free speech protections but also has the potential to strengthen the exercise of that right in the decades ahead.

THE FUTURE OF FREE SPEECH

OCTOBER 2025

www.futurefreespeech.org