THE RIGHT TO LIE WITH AI?

FIRST AMENDMENT CHALLENGES FOR STATE EFFORTS TO CURB FALSE POLITICAL SPEECH USING DEEPFAKES AND SYNTHETIC MEDIA

DAXTON R. "CHIP" STEWART*

JEREMY LITTAU**

Elections are now taking place in the era of widespread, accessible artificial intelligence (AI) tools, and 20 states have passed laws aimed at curbing the spread of false photos, videos, and audio of candidates. The authors review deepfake and AI technology and legislative efforts to regulate them, finding strong First Amendment protection for false political speech stemming from the Supreme Court's decision in U.S. v. Alvarez. Provisions in state laws such as prior restraints, electioneering rules, exemptions for parody and satire, and mandatory disclaimers are reviewed, and the authors find that most face significant First Amendment hurdles. As an alternative, the authors conclude with technological solutions such as digital watermarking by AI companies and labeling by online services facilitating distribution.

* Daxton R. "Chip" Stewart, J.D., Ph.D., LL.M., is a professor in the Department of Journalism at Texas Christian University.

** Jeremy Littau, Ph.D., is an associate professor in the Department of Journalism and Communication at Lehigh University.

INTRODUCTION

With the advent of digital tools that make fake videos, photographs, and audio both appear more real and are easier to use than ever, worries about malicious use of these tools to influence voters have resulted in numerous efforts to limit or ban their use in political campaigns.

Manipulation efforts have come from a variety of sources. The Republican presidential nomination campaign of Florida Gov. Ron DeSantis in 2023 distributed images on Twitter of former President Donald J. Trump hugging and kissing Dr. Anthony Fauci that appeared to be created using artificial intelligence (AI) image generators. The images did not include any disclosures or disclaimers that they were fake, and they were mixed into real photos and videos to make them appear even more real. And while DeSantis campaign staff refused comment, they noted that they were fighting fire with fire because the Trump campaign had been "continuously posting fake images and talking points to smear the governor."[1]

But the deception is not only coming from campaigns. The BBC reported that Trump supporters were using AI-generated photos to target Black voters, generating dozens of such images in 2024 and spreading them across social networks. Conservative radio host Mark Kaye posted one such image, which he created "of Mr. Trump smiling with his arms around a group of black women at a party" and shared to his Facebook audience of more than 1 million on top of an article about Black voters supporting Trump. Kaye said he

1. Alexandra Ulmer & Anna Tong, *With apparently fake photos, DeSantis raises AI ante*, REUTERS, (June 8, 2023, 21:08 MDT), https://www.reuters.com/world/us/is-trump-kissing-fauci-with-apparently-fake-photos-desantis-raises-ai-ante-2023-06-08 [https://perma.cc/S8H6-PLCB].

was not a "photojournalist," nor was he claiming the image was accurate or real, just that he was a "storyteller."[2]

These images would run afoul of many state laws or bills under consideration that extend civil liability, criminal penalties, and even injunctive relief for deepfakes and deceptive use of AI in political campaigns. Texas and California passed the first of these kinds of laws in 2019,[3] and several other states have followed with a variety of limits and remedies available to address the use of AI tools in spreading political misinformation and disinformation. Public Citizen has been tracking these efforts and counts more than 80 bills introduced in state legislatures since 2019.[4] At least 20 have become law, according to the National Conference of State Legislatures (NCSL).[5] Regulations include mandatory labeling that the videos or photos are fake and generated by AI, such as in Michigan,[6] as well as limits on how close to elections deceptive use of AI might be used, such as in Texas, which criminalizes "deep fake videos" distributed in the thirty days before an election.[7] Among the most expansive of these efforts was a bill in Georgia that would have made creating or distributing AI-generated content with intent to manipulate an election a felony punishable by a minimum of two years in prison and up to a $50,000 fine.[8]

The Georgia House, which passed the bill with overwhelming bipartisan support, identified "the rapid increase and use in advancements of artificial intelligence and other sophisticated technologies" that pose "a unique danger to the State of Georgia's free and fair system of elections."[9] The bill is an effort, legislators wrote, to craft a law that recognizes protection of the "utmost rights to both free and fair elections and freedom of speech" and is

2. Marianna Spring, *Trump supporters target black voters with faked AI images*, BBC NEWS (Mar. 4, 2024), https://www-bbc-com.cdn.ampproject.org/c/s/www.bbc.com/news/world-us-canada-68440150.amp [https://perma.cc/P9LM-DTUF].

3. *See* Alexandra Tushman, *"Malicious Deepfakes" – How California's A.B. 730 Tries (and Fails) to Address the Internet's Burgeoning Political Crisis*, 54 LOY. L.A. L. REV. 1391 (2021).

4. *See Tracker: State Legislation on Deepfakes in Elections*, PUBLIC CITIZEN (last updated Apr. 15, 2025), https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/ [https://perma.cc/9EHQ-DLJ8].

5. *See Artificial Intelligence (AI) in Elections and Campaigns*, NAT'L CONF. OF STATE LEGISLATURES (last updated July 23, 2025), https://www.ncsl.org/elections-and-campaigns/artificial-intelligence-ai-in-elections-and-campaigns [https://perma.cc/RJ6V-JBF4].

6. H.B. 5141, 102d Leg., Reg. Sess. (Mich. 2023).

7. TEX. ELEC. CODE ANN. § 255.004(d) (West 2024).

8. H.B. 986, 157th Gen. Assemb., 2d Reg. Sess. (Ga. 2023).; *see also* Mark Niesse, *Georgia House backs ban on election deepfakes*, AJC POLITICS (Feb. 22, 2024), https://www.ajc.com/politics/georgia-house-approves-criminalizing-deceptive-deepfakes-in-elections/ZBLFSHZRDJDXTDLVUIJ45WAK2U [https://perma.cc/4B66-UFLR].

9. H.B. 986, 157th Gen. Assemb., 2d Reg. Sess. (Ga. 2023).

"narrowly tailored for the purpose of protection against the use of deceptive media in bad faith to influence elections."[10]

Opponents in the Georgia Senate painted the bill as an "attack on 'memes' used in political discourse" and an effort to stifle satire.[11] This led to one state senator, a co-sponsor of the bill, to use AI tools to generate audio of two of the bill's more vocal opponents in which their voices are manipulated to say they actually support the bill. The video, posted on YouTube in and presented to the Senate's judiciary committee, includes the following in the manipulated voice of one senate opponent: "I would ask the committee: how is using my biometric data, like my voice and likeness, to create media supporting a policy that I clearly don't agree with the First Amendment right of another person?"[12] Ultimately, the bill died without a vote in the Senate when the 2024 legislative session ended.[13]

While such deepfake and AI political speech laws have been enforced or prosecuted, they are starting to face challenges. A deepfake parody video of Vice President Kamala Harris, created by a user calling himself "Mr. Reagan," in which her voice was emulated to say she was the "ultimate diversity hire" among other insults, circulated widely on X (formerly known as Twitter) in 2024, without any labeling or disclaimers as required by a new round of laws passed by the California legislature.[14] The video creator sued in federal district court and won a preliminary injunction, as the Eastern District of California found that the California law in question was "a hammer instead of a scalpel, serving as a blunt tool that hinders humorous expression and unconstitutionally stifles the free and unfettered exchange of ideas which is so vital to American democratic debate."[15]

As Judge Mendez noted in the ruling, although AI-generated audio, photos and videos may pose risks to free and fair elections, efforts to limit or ban them also bring significant risk of unintended

10. *Id.*

11. George Chidi, *Georgia lawmakers are using an AI deepfake video to try to ban political deepfakes*, GUARDIAN (Mar. 20, 2024, 17:21 EDT), https://www.theguardian.com/us-news/2024/mar/20/georgia-ai-ban-political-campaigns-deepfakes [https://perma.cc/7WAH-KA6V].

12. AI Reagan, *HB 986 w Disclaimer 2*, YOUTUBE, at 1:44-1:57 (Mar. 19, 2024), https://youtu.be/3VPXPA2iUZI?si=2J8PnBtm9MUWuBw4.

13. H.B. 986, 2022-23 GA. GEN. ASSEMB, 2d Sess., https://www.legis.ga.gov/legislation/66172 [https://perma.cc/6KNE-XRC7].

14. Lara Korte, *Creator of Kamala Harris parody video sues California over election 'deepfake' ban*, POLITICO (Sept. 18, 2024, 22:00 EDT), https://www.politico.com/news/2024/09/18/california-deepfake-ban-lawsuit-harris-00179975 [https://perma.cc/RB88-P7SJ].

15. Kohls v. Bonta, 752 F. Supp. 3d 1187, 1199 (E.D. Cal. 2024) (order granting Plaintiff's motion for preliminary injunction).

consequences or chilling effects on political speech[16] – a core protection of the First Amendment in the United States. False speech including even outright lies have broad protection under the First Amendment, and drafting laws to withstand Constitutional scrutiny presents challenges for regulators and lawmakers. And it's possible that legislators and politicians are overreacting to new technology that existing law and other technical solutions may already adequately address.[17]

The purpose of this paper is to examine state laws attempting to limit or ban AI-generated false political speech, how courts may apply the First Amendment to such efforts, and to provide policy guidance to jurists and legislators trying to address these challenges. The authors start with a look at what is new and unique about AI technology that has made this a particular challenge in the context of elections. Then, the authors address First Amendment issues – specifically, protection for algorithmic content generation and protection for false political speech – that present hurdles for regulators who would have to defend the constitutionality of such laws, including a review of common provisions in the laws that have been enacted. Finally, the authors conclude with recommendations on how best to craft legislation or other technological policy to balance the negative effects of election misinformation concerns with robust free speech protection.

I. BACKGROUND

Artificial intelligence is a term that comes from John McCarthy, who worked to create a category of computing dedicated to "the science and engineering of making intelligent machines."[18] AI systems are marked by their ability to autonomously create a path to complete a task using goal and reward system put in place by a human programmer. The path to completion is determined by the AI using a machine-learning system that simulates different ways of solving the problem and attempts to determine the quickest or best path depending on the goal and reward.[19] In other words, a programmer sets the goal but the AI determines the way to accomplish the task based on how it has been trained.

---

16. *Id.*

17. David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, ELECTRONIC FRONTIER FOUND. (Feb. 13, 2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them [https://perma.cc/H3HV-66BB].

18. Kemal Gökham Nalbant, *The Importance of Artificial Intelligence in Education: A Short Review*, J. REV. IN SCI. & ENG'G, 1 (2021).

19. Hongmei He et al., *The Challenges and Opportunities of Human-Centered AI for Trustworthy Robots and Autonomous Systems*, IEEE TRANSACTIONS ON COGNITIVE & DEV. SYS., 1 (2021).

Rather than thinking of AI as a process by which a computer thinks, it's more accurate to describe it in mathematical terms. AI attempts to mimic the processes found in human intelligence by calculating the probabilities for a desired response based on human data and choices. Humans learn through information-seeking and synthesis to envision solutions to a given problem. Information-seeking for humans could come through education or curiosity, but the building block is always some past set of knowledge produced by an external source. For AI, this education is known as "training data" that can teach a computer everything from language syntax to what a given object looks like. The process of synthesis and decision-making are bounded in part by the task; that is, what a person wants to do with learned information constrains the set of choices and thus the output of the process of turning knowledge into action.[20] For an AI system, these boundaries are the types of guardrails many products deploy, such as a prohibition on producing election misinformation or producing sexually explicit imagery.

AI systems have training data to learn about language structures, concepts, and ideas, and they use that data along with data in the form of information libraries to help it predict correct decisions in the form of outputs. What AI produces is the result of mathematical probabilities, and this is crucial to understanding these systems.[21] An AI system is attempting to meet user expectations given the prompt, calculating the probability that an image is in line with what a user wants based on a prompt, or that the next word in a sentence for a text response is likely to be the correct one given all it has learned from the material it is trained on. What AI produces is its best prediction of what a user wants given a particular query.

To build a product that meets user prompt expectations, AI generators need a substantial library of information to draw on.[22] If a human is asked to sketch a picture of an apple from memory, for example, it draws on mental imagery from experiences such as physically touching an apple or viewing one in a photograph to conceptualize the look and feel of the object, but the output could vary from drawing to drawing such that you'll get different colors or sizes. Apples have a "look," but they also take different forms and colors; there is a difference between the concept of the object and a particular artifact a human makes to render that object unless the

20. Meredith Broussard, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (MIT Press, 2021).

21. *Id*.

22. Weixin Liang et al., *Advances, Challenges and Opportunities in Creating Data for Trustworthy AI*, 4 NATURE MACH. INTEL. 669–77 (2022).

request is highly specific (i.e. "a red apple that is more tall than wide, with a stem and a leaf coming out of the top"). AI mimics this human process, but instead of mental imagery, an AI is drawing on training data (large text data sets from books or news sources, or large libraries of photos and video). This data consists of isolated parts of images with labels and tags that taught the AI system what a particular object or scene looks like. Similarly, if you were to ask a chatbot to answer a question about a concept or a news event, its library of material would help it predict the assembly of words likeliest to be the best synthesis of the library of data it is drawing on, and its accuracy would in part depend on the specificity of the question.[23]

A. AI Outputs

This research is interested in two different types of AI outputs, which are the result of a user query (usually a string of text) that indicates the desired output goal. For the purposes of this research, outputs can be classified in two ways: *process* and *product*.

Outputs that are *processes* are defined as AI operating in the background to manage a task without specific human oversight on it from start to finish. Human intervention comes at the programming stage in terms of managing the source data and constructing or altering the variable weights that an AI would use to make decisions. Applying the definition of AI posited earlier, algorithms work on probability models to estimate the best decision based on preset parameters that incorporate "best" decisions from training data and programmed factors intended to shape future decisions.[24] Using machine learning, a process AI makes choices and then learns from those choices to help it improve its decision-making.

Algorithms are a classic example of an AI process, running in the background to make some services more efficient. Algorithms use preset programmed variables and rules to give a computer system a path to follow in the process of decision making, allowing them to mimic human processes but at a far greater scale given the data quantity. Algorithms also are able to adapt to new data and

23. Matthew Burtell & Helen Toner, *The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1*, CTR. FOR SEC. & EMERGING TECH. (Mar. 8, 2024), https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/ [https://perma.cc/5QTJ-35ZT].

24. Kyle Wiggers, *Are AI models doomed to always hallucinate?*, TECHCRUNCH (Sept. 24, 2023, 6:30 AM PDT), https://techcrunch.com/2023/09/04/are-language-models-doomed-to-always-hallucinate/? [https://perma.cc/L7KE-MA4A].

inputs by accounting for the history of user actions to influence future actions over time.[25]

Algorithms are everywhere in our digital lives. For example, logistics specialists such as shipping companies use algorithms to sort and organize orders and deliveries, and online businesses use them to generate custom product recommendations.[26] But digital media companies also use algorithms. Social networking platforms such as Facebook use algorithms to create a customized news feed that gives users a landing page tailored to their own interests.[27] Search engines such as Google use algorithms to learn from past searches, using data acquired from search engine use to improve results going forward.[28] Many have theorized that the future of online information search will look more like an AI chatbot than a traditional Google search result; both processes use algorithms to rank and sort information, but Microsoft offers a case study of where things are going. Its Copilot AI chatbot, built on OpenAI's protocol, delivers specific generative AI text answers synthesized from billions of documents. In contrast, search engines such as Bing or Google provide lists of links without context or analysis[29] and offload the process of finding answers onto users who have to sift through several sources and come to their own conclusions. There is a different emphasis on how you find answers, and AI is a more passive search activity than scrolling through search results and finding information for yourself.

One thing to note about algorithms is that they are the result of weights that reflect choices made by programmers about what is most relevant or valuable. A useful way to see this is by examining the changes made by X, the company formerly known as Twitter. Before Elon Musk purchased Twitter, the platform used algorithmic moderation as a type of shield; the algorithm predicted whether specific material would be perceived as harmful by average

25. Daisuke Wakabayashi, *Google Dominates Thanks to an Unrivaled View of the Web*, N.Y. TIMES (Dec. 14, 2020), https://www.nytimes.com/2020/12/14/technology/how-google-dominates.html [https://perma.cc/669R-LKLU].

26. Nicholas Shields, *UPS is turning to predictive analytics,* BUS. INSIDER (July 20, 2018, 8:22 AM MT), https://www.businessinsider.com/ups-using-predictive-analytics-algorithm-2018-7 [https://perma.cc/E5AW-K7VB].

27. Mike Isaac & Sheera Frankel, *Facebook's Algorithm Is 'Influential' but Doesn't Necessarily Change Beliefs, Researchers Say*, N.Y. TIMES (July 27, 2023), https://www.nytimes.com/2023/07/27/technology/facebook-instagram-algorithms.html [https://perma.cc/R4RY-ZGMY].

28. Wakabayashi, *supra* note 25.

29. Tom Warren, *Microsoft's next big AI push is here after a year of bing*, VERGE (Feb. 7, 2024, 8:00 AM MST), https://www.theverge.com/2024/2/7/24064440/microsoft-super-bowl-ad-ai-copilot [https://perma.cc/T5XP-LDX6]; *See also* Matt Honan, *AI means the end of internet search as we've known it*, MIT TECH. REV. (Jan. 6, 2025) https://www.technologyreview.com/2025/01/06/1108679/ai-generative-search-internet-breakthroughs/ [https://perma.cc/R7PJ-U25Y].

users and then labeled them with warnings or filters that would get activated for users if they had opted into having content labeled.[30] Musk made two changes that had a critical impact on the algorithm. First, he got rid of the standard verification badge Twitter had formerly used to designate individuals who were known for specific expertise and could be verified as the person running a particular account that was given that badge. In place of the old system,[31] Musk implemented a system that let anyone be "verified" for $8 a month. That "verified" status was given extra weight in the algorithm, such that those users who paid were more likely to surface in algorithmic feed decisions and replies on individual posts.[32] As a result of policy changes, the algorithm changed, and many saw the user experience degraded as verified experts were put on the same priority level in algorithmic rankings as trolls and troublemakers who were willing to pay X a monthly fee.[33]

Algorithms are used in other ways in media products, such as news sites using them to create a customized experience similar to a social platform's news feed (with the key distinction that it's digital gatekeeping of professionally produced news content done by automation rather than human editors). Social platforms also use algorithms to moderate content, allowing them to avoid expensive human moderator costs or subjecting human moderators to some of the most disturbing content on the platform, such as violent extremist content or child sexual abuse material. Some platforms also use a hybrid model that lets algorithms handle some of the more basic decisions, while humans act as a second check on

30. *See, e.g.*, Casey Newton, *The Trauma Floor*, VERGE (Feb. 25, 2019, 6:00 AM MST), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/24E2-R8YR]; Copia Institute, *Content Moderation Case Study: Twitter's Algorithm Misidentifies Harmless Tweet As 'Sensitive Content' (April 2018)* TECHDIRT (Sept. 25, 2020, 15:30 MDT), https://www.techdirt.com/2020/09/25/content-moderation-case-study-twitters-algorithm-misidentifies-harmless-tweet-as-sensitive-content-april-2018/ [https://perma.cc/Y47W-UY76].

31. Jon Porter, *Twitter begins removing blue checkmarks from all legacy users*, VERGE (Apr. 20, 2023, 12:16 MDT), https://www.theverge.com/2023/4/20/23690820/twitter-verified-blue-checkmark-legacy-elon-musk [https://perma.cc/RG7B-RWVS].

32. James Vincent, *Twitter says paying blue subscribers now get 'prioritized rankings in conversations',* VERGE (Dec. 23, 2022, 4:13 AM MST), https://www.theverge.com/2022/12/23/23523845/twitter-blue-paying-priority-replies-conversations [https://perma.cc/7GB8-XKWX].

33. Amanda Yeo, *Twitter's made 'legacy' verified blue ticks indistinguishable from paid ones*, MASHABLE (Apr. 3, 2023), https://mashable.com/article/twitter-blue-tick-legacy-description-april-1 [https://perma.cc/WY38-7CZQ].

the system if the issue is complicated or a user objects to a decision made by an automated process.[34]

Outputs we define as *products* build on process AI outputs. They are often referred to in public conversation as "generative AI," products that use the algorithms found in process AI to create something original and new based on a text query. Thus, they provide a tangible result where the user both inputs the request and is given a result not akin to a search result, but rather synthesizing a new output from its training data and information libraries.[35] Generative AI is often the domain of public-facing tools that allow a user to use a prompt to get an AI to make something specific and offer follow-up prompts to shape the result further. These types of tools output a result to a text prompt (sometimes using an uploaded audio file, image, or video as a reference) to create something synthetic and original. Many AI products are digital media outputs that create original text via a large-language model (LLM), audio, images, or video.[36] Examples include a JPG image created from an AI image generator such as OpenAI's DALL-E, the Midjourney generative AI project on Discord, or an answer in text query from a prompt given to a chatbot interface such as ChatGPT or Anthropic. AI products such as RunwayML have a suite of tools for photo, video, and audio creation and editing, allowing you to upload content that builds on concepts of a face or a voice and then transform that base material into something new by typing in a text description of what you want. Adobe also has an AI tool embedded in its Photoshop product. AI products can also be combined. If a person wants to create an AI message, they could generate an AI image on DALL-E and combine that with generative text produced by a chatbot to create entirely new and believable messages for circulation.

The quality of the library can lead to multiple problems. An incomplete or low-quality data set could lead to bad outputs that reinforce structural problems such as misogyny or racist content. This "garbage in, garbage out" phenomenon reinforces the need for careful curation of an AI's source material lest it build on the ignorance and hatred reflected in some historical media content or

34. Tomas Apodaca & Natasha Uzcategui-Liggett, *How Automated Content Moderation Works (Even When It Doesn't)*, MARKUP (Mar. 1, 2024, 8:00 AM UTC), https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work [https://perma.cc/Q4W8-FN58].

35. Kevin Roose & Cade Metz, *How to Become an Expert on A.I.*, N.Y. TIMES (Mar. 31, 2023), https://www.nytimes.com/article/ai-artificial-intelligence-chatbot.html [https://perma.cc/F3S6-AVTD].

36. Adam Zewe, *Explained: Generative AI*, MIT NEWS (Nov. 9, 2023), https://news.mit.edu/2023/explained-generative-ai-1109 [https://perma.cc/SA68-CFBR] (providing an overview of AI product outputs).

even the abundance of low-quality, self-published content in the Internet age.[37] But the reverse also comes with problems; better training data could lead to fake content that is indistinguishable from something made by human hands, in turn giving creators the ability to pass off these outputs as real.[38] It is the latter images and videos that are troubling legislators and policymakers who fear effects on voters and elections.

B. Generative AI and Media

The platforms used to generate these media artifacts are not necessarily distribution outlets, so using anything created requires a second step of a user repurposing the content to a platform, such as uploading an AI product to a social platform like X or posting it on a website. This has created content moderation problems for social platforms with regard to realistic-looking AI content that is false, misleading, or damaging. For example, in early 2024, X had to temporarily ban all searches for Taylor Swift's name because a collection of bots was distributing fake pornographic AI images of Swift.[39] Meta, concerned about misinformation and election integrity on its Facebook and Instagram platforms, has also encountered this problem with generative AI images. As a result, it says it is moving to more tightly monitor and label generative AI imagery ahead of the 2024 U.S. elections.[40]

Beyond the platforms, there are situations when AI content posted online breaks into the news or it is used for purposes that becomes the basis for breaking news. In late 2023, an AI audio recording of a candidate purportedly talking about rigging election

37. *See, e.g.*, Sigal Samuel, *Black Nazis? A woman pope? That's just the start of Google's AI problem*, VOX (Feb. 28, 2024, 5:30 MST), https://www.vox.com/future-perfect/2024/2/28/24083814/google-gemini-ai-bias-ethics [https://perma.cc/4UH3-XNK6]; Reed Albergotti, *Adobe Firefly repeats the same AI blunders as Google Gemini*, SEMAFOR (Mar. 13, 2024, 12:13 MDT), https://www.semafor.com/article/03/13/2024/adobe-firefly-repeats-the-same-ai-blunders-as-google-gemini [https://perma.cc/867P-WPZK].

38. Huo Jingnan, *AI-generated text is hard to spot. It could play a big role in the 2024 campaign*, NPR (June 29, 2023, 5:00 AM ET), https://www.npr.org/2023/06/29/1183684732/ai-generated-text-is-hard-to-spot-it-could-play-a-big-role-in-the-2024-campaign [https://perma.cc/37R8-U7V7].

39. Mallory Moench, *Taylor Swift Searches Blocked by X Amid Circulation of Deepfakes*, TIME (Jan. 28, 2024, 14:53 EST), https://time.com/6589487/taylor-swift-searches-blocked-x-twitter-deepfakes-response/ [https://perma.cc/E2J9-JLS2].

40. Hayden Field, *Meta says it will identify more AI-generated images ahead of upcoming elections*, CNBC (Feb. 6, 2024, 14:29 EST), https://www.cnbc.com/2024/02/06/meta-to-identify-more-ai-generated-images-ahead-of-upcoming-elections.html [https://perma.cc/ELR2-F3X4].

results circulated in the days before votes were cast in Slovakia.[41] In the U.S., AI audio mimicking Joe Biden's voice encouraging people not to vote in the Democratic Party primary was circulated by an automated robocall in the days before the New Hampshire election.[42] Some experts have declared that 2024 will be the year of the AI election because of the low-cost or free AI tools that are available to the general public.[43]

Of course, fake media images used to influence politics in particular are not a new phenomenon. Editing a photo to change the meaning of the picture has been considered such a problem for institutions charged with truth-telling that organizations like the National Press Photographers Association have banned the practice.[44] But in the arena of governance, the existence of guardrails is nearly entirely up to public sentiment and a rigorous press accountability structure. During Joseph Stalin's reign in the Soviet Union, his government routinely manipulated photos using techniques that were sophisticated at the time. Stalin's team often removed people from photos purportedly documenting historical events either because they became enemies of the state or because association with Stalin was bad public relations in hindsight.[45]

It's notable that these techniques or even more sophisticated digital editing in Photoshop are more in line with deepfakes, which are a specific subset of misinformation content. Deepfakes are defined not by original generation, as with AI image creators. Rather, deepfake refers to "a specific kind of synthetic media where a person in an image or video is swapped with another person's likeness." In other words, the term refers to the use of existing media artifacts to combine them into a synthesis of the originals. The term comes from the name of a Reddit user who "created a space on the online news and aggregation site, where they shared pornographic videos that used open-source face-swapping

41. Curt Devine et al., *A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning*, CNN (Feb. 1, 2024, 6:09 AM EST), https://www.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html [https://perma.cc/JG9N-WSM5].

42. Em Steck & Andrew Kaczynski, *Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's Democratic primary*, CNN (Jan. 22, 2024, 17:44 EST), https://www.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html [https://perma.cc/53RS-6VWP].

43. Jingnan, *supra* note 38.

44. *See National Press Photographers Association: Code of Ethics*, NEWS LEADERS ASS'N, https://members.newsleaders.org/resources-ethics-nppa [https://perma.cc/P94Y-VYAB].

45. Erin Blakemore, *How Photos Became a Weapon in Stalin's Great Purge*, HISTORY (Apr. 20, 2018), https://www.history.com/articles/josef-stalin-great-purge-photo-retouching [https://perma.cc/W6KZ-Y2SD].

technology."[46] So, using the DALL-E image generator to create an image of Joe Biden robbing a bank would not be a deepfake. Instead, a deepfake could involve someone taking an image of a person robbing a bank, cutting out Biden's head from a different image, and then combining the two using hand tools or a photo editor such as Adobe Photoshop. Both deepfakes and generative AI images represent a larger universe of content known as "synthetic media"–a phrase used by some states in their laws targeting false political speech[47] –but they are differentiated by the technique used to create them. AI can be created from a simple word prompt that harnesses billions of images in a library to create something original and thus untraceable.[48] In that sense, all deepfakes are synthetic media, but not all synthetic media (which includes AI) are deepfakes. Still, it's important to note that in the public conversation and media coverage, AI fakes are often referred to as "deepfakes" in the news–this happened routinely in the Taylor Swift incident on X–which complicates the public and regulatory conversation about them. It is possible that, over time, AI images will simply become another type of deepfake in common understanding. Thus, the act of construction is a key differentiator worth remembering.

Media manipulation doesn't mean altering words or images wholesale. Basic video editing techniques, such as editing cuts or changing the speed of the video or audio clip, can have the effect of changing meaning as well. This was the case in 2019, when an altered clip of then House Speaker Nancy Pelosi began circulating online to give the appearance her speech was slurred. It was even shared by then-President Donald Trump as an allegation of her being drunk in public.[49] The edited Pelosi video was dubbed a "cheapfake" by some observers.[50] Deepfakes and generative AI share the reality that mere creation of an artifact does not itself guarantee impact. Any sort of synthetic media built for disinformation or propaganda relies on a distribution channel, such as the news in the case of the Soviets or a social media platform in

46. Meredith Somers, *Deepfakes, explained*, MIT SLOAN (July 21, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained [https://perma.cc/NEF3-NPNR].

47. *See, e.g.*, IDAHO CODE § 67-6628A(2)(c) (2024); UTAH CODE ANN. § 20A-11-1104(1) (LexisNexis 2024); WASH. REV. CODE § 42.62.020(1) (2024).

48. Jake Traylor, *AI-generated 'synthetic media' is starting to permeate the internet*, NBC NEWS (Mar. 3, 2023, 12:00 MST), https://www.nbcnews.com/tech/tech-news/ai-generated-synthetic-media-future-content-rcna72958 [https://perma.cc/Z9XC-T5PT].

49. Joan Donovan & Britt Paris, *Beware the Cheapfakes*, SLATE (June 12, 2019, 17:57), https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html [https://perma.cc/B6W5-E5ZV].

50. Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 15–16 (2020).

more recent times. Social platforms have an additional layer of challenge compared to news because every user is a potential spreader and amplifier.

The ability of false political information to spread rapidly has drawn the attention of legislators, who have targeted AI-generated audio, video, and photographs with regulations aimed at curbing their potential influence on voters and elections.

To understand the extent to which these AI tools may be regulated in the context of political campaigns, the authors set out to examine the three areas of law. We begin by considering the extent to which the algorithmic outputs of machines are protected as "speech" by the First Amendment. Next, considering the innovations in AI fueling current image, video, and audio generation tools, we explore how First Amendment protections apply to these tools, particularly regarding matters of false political speech. Finally, we consider how courts may apply the First Amendment to current laws and bills aiming to ban or limit AI-generated and/or deepfake images, video, and audio in the context of political misinformation or disinformation, specifically addressing common features in state laws targeting political deepfakes.

II. ANALYSIS

A. Algorithmic Outputs as "Speech"

Scholars and jurists have been examining possible free speech rights for chatbots and AI for about a decade. Early efforts examined protections for outputs of search engines such as Google, whose outputs were compared by Eugene Volokh and Donald Falk to the editorial judgment of websites and newspapers. Volokh and Falk called these outputs "human editorial judgments (that) are responsible for producing the speech displayed by a search engine," thus deserving protection as "the speech of the corporation, much as the speech created or selected by corporate newspaper employees is the speech of the newspaper corporation."[51] In 2014, a federal district judge cited the Volokh and Falk article in dismissing a case against the Chinese search engine Baidu on First Amendment grounds.[52] The court found "a strong argument to be made that the First Amendment fully immunizes search-engine results from most, if not all, kinds of civil liability and government regulation." Their holding was rooted in the general rule that the First

51. Eugene Volokh & Donald M. Falk, *Google First Amendment Protection for Search Engine Search Results*, 8 J.L. ECON. & POL'Y 884, 888–89 (2012).

52. Zhang v. Baidu.com, 10 F. Supp. 3d 433, 436 (S.D.N.Y. 2014).

Amendment does not allow the government to interfere with the editorial judgments of speakers, and that search engines "inevitably make editorial judgments about what information (or kinds of information) to include in the results and how and where to display that information."[53]

One of the first situations examining AI activity with speech elements involves the facial recognition company ClearviewAI (Clearview), which has faced numerous lawsuits for its AI-powered tools that scrape data (including photographs) posted online and uses it to build individual profiles, that are sold to companies and law enforcement.[54] Clearview argued that "the capture of faceprints from public images and (their) analysis of the public faceprints is protected speech," thus making enforcement of the Illinois Biometric Information Privacy Act (BIPA) a violation of its First Amendment rights.[55] A federal district court found, however, that Clearview's process "involves both speech and nonspeech elements," triggering an intermediate scrutiny analysis. Which was at least enough to overcome Clearview's motion to dismiss because Illinois had an important interest in protecting citizens' biometric data and because BIPA was not, at least on its face, broader than necessary to serve that interest.[56]

Other lawsuits ostensibly rooted in the First Amendment attempting to hold social media sites such as YouTube[57] and Facebook[58] responsible for the acts of their algorithms have not turned on whether those outputs were a kind of speech, rather on the nature of the companies employing the algorithm. These are private companies, not state actors, and courts have found that the First Amendment does not provide a remedy for individuals complaining about the effects of the algorithm. Indeed, social media companies have a "First Amendment right to decide what to publish and what not to publish" on their platforms,[59] including the output of their algorithms to restrict or limit content posted by other private speakers. The content moderation decisions of platforms,

53. *Id.* at 438.

54. Jonathan Stempel, *Face scanner firm Clearview AI agrees to limits to settle lawsuit*, REUTERS (May 9, 2022, 14:02 MDT), https://www.reuters.com/technology/face-scanner-firm-clearview-ai-agrees-limits-settle-lawsuit-2022-05-09/ [https://perma.cc/R7FK-7K7R].

55. *In re* Clearview AI, Inc., Consumer Priv. Litig., 585 F. Supp. 3d 1111, 1120 (N.D. Ill. 2022).

56. *Id.* at 1121.

57. *See* Newman v. Google, 687 F. Supp. 3d 863 (N.D. Cal. 2023); Prager Univ. v. Google LLC, 301 Cal. Rptr. 3d 836, 849 (Cal. Ct. App. Sixth Dist. 2022); Prager Univ. v. Google LLC, 951 F.3d 991 (9th Cir. 2020).

58. *See* Children's Health Def. v. Facebook, Inc., 546 F. Supp. 3d 909 (N.D. Cal. 2021).

59. La'Tiejira v. Facebook, Inc., 272 F. Supp. 3d 981, 991 (S.D. Tex. 2017).

many of which are made by algorithms, "constitute protected exercises of editorial judgment" by the platforms, as recognized by the U.S. Court of Appeals for the Eleventh Circuit in *NetChoice v. Moody* in 2023: "Put simply, with minor exceptions, the government can't tell a private person or entity what to say or how to say it."[60] And although the Supreme Court rejected a facial First Amendment challenge to Florida and Texas laws in the *NetChoice* cases on appeal, Justice Kagan, writing for the majority, found that algorithmic prioritization of content was a form of speech drawing First Amendment protection, noting that "Texas's law profoundly alters the platform's choices about the views they will, and will not convey. And time we have time and again held that type of regulation to interfere with protected speech."[61] Unless the Court changes direction as the *NetChoice* cases are reconsidered on remand, platforms and the output of their algorithms will continue to hold robust First Amendment protections, curbing at least somewhat the ability of government to ban or restrict those outputs.

Ultimately, as detailed in the previous section, what we identify as "AI" today is actually algorithmic output of programs trained on underlying information, such as LLMs or art generators. And courts have repeatedly found that algorithmic output receives at least some First Amendment protection. So far, the words and works are not inspired by a creative, independent being; rather, they are the result of human programming, with outputs generated from other previously existing works (whether human or AI-generated). The U.S. Copyright Office recognized this with several rulings in 2023 rejecting efforts to register artistic outputs aided or entirely generated by AI tools because they do not "contain sufficient human authorship necessary to sustain a claim to copyright."[62] We are not yet to questions about whether "strong AI"–that is, machines capable of independent thought and creation, popular in science fiction depictions of malicious computers and robots–are deserving of First Amendment protection for their speech.[63] But scholars considering that very point have concluded that even these potentially more advanced forms of AI would have some First Amendment protections afforded to their outputs. As just one example, Toni Massaro, Helen Norton and Margot

60. NetChoice, LLC v. Att'y Gen. of Fla., 34 F.4th 1196, 1203 (11th Cir. 2022).

61. Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2391 (2024).

62. Second Request for Reconsideration for Refusal to Register SURYAST (SR # 1-11016599571; Correspondence ID: 1-5PR2XKJ), U.S. Copyright Off. Rev. Bd. (Dec. 11, 2023), https://copyright.gov/rulings-filings/review-board/docs/SURYAST.pdf [https://perma.cc/V5PV-7ZBU].

63. Toni M. Massaro, Helen Norton & Margot Kaminski, *Siri-ously 2.0: What Artificial Intelligence Reveals About the First Amendment*, 101 Minn. L. Rev. 2481, 2942 (2018).

Kaminski identified a variety of First Amendment values and theories, both positive (speech providing value to listeners) and negative (curbing governmental control of speech) that would undergird legal support for speech by "strong AI."[64] As they wrote, "the United States Supreme Court now emphasizes listeners' interests in free speech outputs—rather than speakers' humanness or humanity—in ways that make it exceedingly difficult to place AI speakers beyond the First Amendment's reach."[65]

Thus, the First Amendment does seem to apply to algorithmic outputs, recognizing them as a kind of speech or expression deserving of some protection. We proceed to address exactly what kinds of protection may extend to AI outputs or the people using them, especially when they are being used to spread false political information.

B. First Amendment Protection for False Political Speech

If algorithmic outputs qualify as speech deserving of at least some First Amendment protection, as established supra, then regulation of algorithmic outputs in the context of false political speech faces major hurdles in First Amendment doctrine.

As a starting point, political speech–the target of laws limiting or banning deepfakes in campaigns–receives the highest level of protection possible under the First Amendment. As the Supreme Court reasoned in *Monitor Patriot Co. v. Roy* in 1971, "it can hardly be doubted that the constitutional guarantee has its fullest and most urgent application precisely to the conduct of campaigns for political office."[66]

The centrality of political speech to the First Amendment–recognized as what scholars sometimes call "high-value speech"–makes efforts to curb that speech difficult. High-value speech is recognized as such because of its "contribution to the First Amendment's core functions," including "promotion of democratic self-governance."[67] Political speech is at "the core of the protection afforded by the First Amendment," as the Supreme Court noted in 1995 when it struck down an Ohio law banning distribution of anonymous leaflets and handbills.[68] More recently in the *Citizens United* case when the Supreme Court struck down campaign finance laws limiting spending on campaigns by corporations and

64. *Id.* at 2483.

65. *Id.*

66. Monitor Patriot Co. v. Roy, 401 U.S. 265, 272 (1971).

67. Alan K. Chen & Justin Marceau, *High Value Lies, Ugly Truths, and the First Amendment*, 68 VAND. L. REV. 1435, 1441 (2015).

68. McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 346 (1995).

unions, the Court identified political speech as "central to the meaning and purpose of the First Amendment."[69]

The Court has applied this philosophy in a number of contexts over the years, such as requiring a heightened standard of proof for public officials seeking damages or criminal penalties for defamation, including a showing of knowing or reckless falsehood.[70] The same heightened standard of proof for incitement and true threats was reaffirmed by the Court in *Counterman v. Colorado* in 2023, as Justice Kagan recognized fears that "efforts to prosecute incitement would not bleed over, either directly or through a chilling effect, to dissenting political speech at the First Amendment's core."[71]

For decades, nevertheless, states have passed and attempted to enforce various laws targeting false political speech, and First Amendment challenges to them have had mixed success. A federal district court, in a decision upheld without opinion by the U.S. Supreme Court, struck down portions of New York's Fair Campaign Code in 1976. The law had prohibited misrepresentation of any candidates' "qualifications," "position," or "party affiliation," but the court found these to be impermissibly overbroad. Even though "deliberate calculated falsehoods when used by political candidates can lead to public cynicism and apathy toward the electoral process," the court reasoned, the state could not tamper "with what it will permit the citizen to see and hear."[72] On the other hand, in 1986, a Michigan state appeals court upheld a lower court injunction against a campaign that had been falsely suggesting its candidate was the incumbent judge, violating the state's "false designation of incumbency" law, but it did not address First Amendment arguments, nor was it appealed further.[73] When the Supreme Court decided *McIntyre* in 1995, striking down the provision against anonymous leaflets, it mentioned but did not directly consider the constitutionality of other Ohio laws that contained "detailed and specific prohibitions against making or disseminating false statements during political campaigns," citing them only as less restrictive means than the anonymous speech ban.[74]

But when the Supreme Court decided *U.S. v. Alvarez* in 2011, the First Amendment's application to false speech shifted. By a 6-3 vote, the Court struck down the federal Stolen Valor Act, a law that

69. Citizens United v. FEC, 558 U.S. 310, 329 (2010).

70. *See* N.Y. Times v. Sullivan, 376 U.S. 254, 280 (1964); Garrison v. Louisiana, 379 U.S. 64, 67 (1964).

71. Counterman v. Colorado, 143 S. Ct. 2106, 2118 (2023).

72. Vanasco v. Schwartz, 401 F. Supp. 87, 101 (S.D.N.Y. 1976).

73. Treasurer of Comm. to Elect Lostracco v. Fox, 150 Mich. App. 617 (1986).

74. *McIntyre*, 514 U.S. at 349.

had criminalized lying about having earned military honors.[75] Alvarez, who had falsely claimed he had been awarded the Congressional Medal of Honor while speaking at a public hearing of a water district board, was indicted for his lies.[76] The Supreme Court found the Stolen Valor Act to be insufficiently protective of First Amendment rights.

The plurality opinion – authored by Justice Kennedy and joined by Chief Justice Roberts, Justice Ginsburg, and Justice Sotomayor – declined to add false speech to other historical categories of unprotected speech such as fighting words, obscenity, true threats, defamation, and fraud. Instead, Kennedy said, what matters is that the speech causes a "legally cognizable harm" associated with the false statement. The Stolen Valor Act was struck down because "[it] targets falsity and nothing more."[77]

The plurality distinguished other permissible restrictions on lying – false statements to government officials, perjury, and bans on falsely claiming to speak on behalf of the government – because they come with recognizable harms such as interfering with the administration of justice or causing financial or property loss.[78] The Stolen Valor Act had no such identifiable harm. Further, it was overbroad – covering all false speech about military honors, whether private or public.[79] And, the plurality determined, regardless of the government's compelling interest in honoring military service, there were better options for achieving that outcome than burdening speech. The government, Kennedy wrote, was unable to demonstrate why counterspeech was not a narrower and better avenue for the government to achieve its desired ends: "The remedy for speech that is false is speech that is true. This is the ordinary course in a free society. The response to the unreasoned is the rational; to the uninformed, the enlightened; to the straight-out lie, the simple truth."[80] Kennedy noted, for example, that when Alvarez's lies at the meeting were discovered, he was widely recognized as a "phony," his lies were reported by the press, and he was "ridiculed online."[81]

The concurrence by Justice Breyer, joined by Justice Kagan, also found the law to be unconstitutionally overbroad, but with important reasoning in the context of political speech, where

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

76. *See* Jeff Kosseff, LIAR IN A CROWDED THEATER: FREEDOM OF SPEECH IN A WORLD OF MISINFORMATION 21–22 (2023) (full and detailed account of the events leading to Alvarez's prosecution).

77. *Alvarez*, 567 U.S. at 719.

78. *Id.* at 720–21.

79. *Id.* at 723.

80. *Id.* at 727 (citing Whitney v. California, 274 U.S. 357 (1927)).

81. *Id.*

citizens should have more "breathing room" for error. "[T]he threat of criminal prosecution for making a false statement can inhibit the speaker from making true statements, thereby 'chilling' a kind of speech that lies at the First Amendment's heart," Breyer noted.[82] In the political speech context, the balance is particularly difficult. While a false political statement may deceive voters into voting for the liar, criminal prosecution of such potential lies is also "particularly dangerous (say, by radically changing a potential election result)," leading to even further censorship of speakers and ideas.[83] Like the plurality, Breyer favored counterspeech as the best remedy for false speech.[84]

Since *Alvarez* was decided in 2011, courts have applied it in a variety of areas where people tried to argue that their lies were protected by the First Amendment. Courts have upheld convictions for defendants' violations of laws against impersonating government officials, such as a U.S. Marshal,[85] a law enforcement officer,[86] and a member of Congress,[87] finding that *Alvarez* did not bar application of these laws. In 2023, the District Court for the District of Columbia rejected former President Trump's First Amendment arguments under *Alvarez* in an effort to dismiss federal indictments for Conspiracy to Defraud the United States, Conspiracy to Obstruct an Official Proceeding, Obstruction of and Attempt to Obstruct an Official Proceeding, and Conspiracy Against Rights. While the president argued he was engaging in protected political speech by spreading election lies and encouraging states to submit false slates of electors, the court ruled that "speech in furtherance of criminal conduct does not receive *any* First Amendment protection" (emphasis added). *Alvarez*, the court reasoned, did not undermine this settled precedent.[88]

Political speech regulations in other contexts, however, have fared poorly under the *Alvarez* analysis. The U.S. Courts of Appeals for the Sixth Circuit and the Eighth Circuit, as well as the Massachusetts Supreme Court, have struck down false political speech laws over the past decade. The logic from each is consistent, stemming from the *Alvarez* decision and finding significant constitutional flaws in the respective state laws under consideration.

---

82. *Id.* at 733 (Breyer, J., concurring) (citing Gertz v. Robert Welch, 418 U.S. 323 (1974)).

83. *Id.* at 738.

84. *Id.*

85. United States v. Bonin, 923 F.3d 523 (7th Cir. 2019).

86. United States v. Chappell, 691 F.3d 388 (4th Cir. 2012).

87. United States v. Tomsha-Miguel, 766 F.3d 1041 (9th Cir. 2014).

88. United States v. Trump, 704 F. Supp. 3d 196, 222 (D.D.C. 2023).

The first of these cases to reach a final ruling was *281 Core Committee v. Arneson*, in which the Eighth Circuit struck down a portion of the Minnesota Fair Campaign Practices Act (FCPA) criminalizing the creation or circulation of false political advertising designed to "promote or defeat a ballot question" when the person "knows is false" or "communicates to others with reckless disregard of whether it is false."[89] Violators faced civil penalties of up to $5,000 as well as referral to county attorneys for prosecution. A pair of grassroots organizations challenged the law on its face, arguing that their political speech was chilled and they faced a credible threat of prosecution. As an opening matter, the Eighth Circuit panel recognized that while *Alvarez* decision addressed false speech generally, "it did not deal with legislation regulating false *political* speech" (emphasis added) which triggers review under the strict scrutiny standard.[90] Strict scrutiny requires the government both to demonstrate a compelling state interest and to establish that the regulation is narrowly tailored to advance that interest. The Eighth Circuit said that while the state may have a compelling interest in preserving "fair and honest elections," the FCPA failed in every possible way on the second part of the test: "no amount of narrow tailoring succeeds because § 211B.06 is not necessary, is simultaneously overbroad and underinclusive, and is not the least restrictive means of achieving any stated goal."[91]

The court found that the mechanism for enforcement – in which anyone could file a complaint at any time, triggering a referral to the state's Office of Administrative Hearings (OAH) – was practically problematic. Ballot measure opponents could tactically file complaints that they trigger review late in a campaign, causing irreparable harm: "For all practical purposes, the real potential damage is done at the time a complaint is filed, no matter the possibility of criminal prosecution down the line … Even before a probable cause hearing, the allegation of the falsity itself likely makes the news circuit and creates a stir in the ongoing political discourse."[92] By creating a process that could be gamed by meritless complaints, the state "only opens the door to more fraud" and "opens a Pandora's box to disingenuous politicking itself."[93]

The least restrictive means to advance the state's interest in fair elections, the Eighth Circuit said, is counterspeech. Citing the plurality opinion in *Alvarez*, the court said: "Possibly there is no greater arena wherein counterspeech is at its most effective. It is

89. MINN. STAT. § 211B.06 (2024).
90. 281 Care Comm. v. Arneson, 766 F.3d 774, 783 (8th Cir. 2014).
91. *Id.* at 785.
92. *Id.* at 792.
93. *Id.* at 796.

the most immediate remedy to an allegation of falsity."[94] Ultimately, the court said, it is up to citizens and not the government to be "the monitor of falseness in the political arena."[95]

The Massachusetts Supreme Court reached the same conclusion using similar logic rooted in *Alvarez* when it struck down the state's political false-statements law in *Commonwealth v. Lucas* in 2015.[96] The law, which had been in place since 1922, barred making or publishing "any false statement in relation to any candidate for nomination or election to public office, which is designed or tends to aid or to injure or defeat such candidate," with a penalty of up to six months in prison and a $1,000 fine.[97] Melissa Lucas, a candidate for state representative, sent out brochures with statements such as "Brian Mannai chose convicted felons over the safety of our families" and "Lawyer Brian Mannai has earned nearly $140,000 of our tax dollars to represent criminals." Mannai, her opponent in the state representative race, filed an application for a criminal complaint under the false-statements law, and "held a press conference announcing the filing … suggesting that the brochures 'could put her behind bars.'"[98] Lucas argued under *Alvarez* that her brochures were protected political speech and that the statute was unconstitutional on its face. Applying strict scrutiny as the appropriate level of review, the Massachusetts Supreme Court found that the state had a "compelling interest in the maintenance of free and fair elections," but that the state had not established that the law was "actually necessary" to advance those interests.[99] Echoing the Eighth Circuit's decision in *281 Care Committee*, the court found that the proper remedy to false political statements is counterspeech: "*Alvarez* teaches that the criminalization of such falsehoods is unnecessary because a remedy already exists: 'the simple truth.'"[100]

Likewise, the court noted similar practical challenges with the law as the Eighth Circuit had, recognizing that the process could be gamed by candidates filing an "unmeritorious application." These applications could not be plausibly adjudicated before votes are cast, as was the situation at hand, when Mannai requested a criminal investigation two weeks before the election.[101] The fact that any citizen could file such a complaint weighed against the constitutionality of the law. These tactics by candidates or their

94. *Id.* at 793.
95. *Id.* at 796.
96. Commonwealth v. Lucas, 472 Mass. 387 (2015).
97. MASS. GEN. LAWS ch. 56, § 42 (2024).
98. See *Lucas*, 472 Mass. at 388–89.
99. *Id.* at 398.
100. *Id.* at 399.
101. *Id.* at 404.

supporters could "divert the attention of an entire campaign from the meritorious task at hand of supporting or defeating a ballot question [or candidate]."[102] As such, the court ruled that the law casts "an unacceptable chill on core political speech" and could not survive strict scrutiny.[103]

The Sixth Circuit, in its 2016 opinion in *Susan B. Anthony List v. Driehaus*, also applied *Alvarez* and mirrored the reasoning of the Eighth Circuit and the Massachusetts Supreme Court in addressing Ohio's false-statements law.[104] The SBA List issued a press release saying that Steven Driehaus, a candidate seeking reelection to the U.S. House, had supported "taxpayer-funded abortions" by voting for the Affordable Care Act.[105] Ohio law limited false statements in a variety of areas about campaigns, including a candidate's voting record, "knowing the same to be false or with reckless disregard of whether it was false or not, if the statement is designed to promote the election, nomination, or defeat of the candidate."[106] First-time violators could be sentenced to prison for up to six months and pay a $5,000 fine. Complaints could be filed by citizens or public officials with the Ohio Elections Commission, and in this case, Driehaus had filed the complaint as a candidate. The law had been in effect since the 1970s, and although the Supreme Court left it untouched without comment in *McIntyre* in 1995, the Sixth Circuit found that any previous decisions upholding the law had been abrogated by the Supreme Court's decision in *Alvarez*.[107]

Because "Ohio's false-statements laws target speech at the core of First Amendment protections—political speech," the Sixth Circuit used strict scrutiny as the standard to review the law.[108] The court ultimately found that Ohio's interest in "preserving the integrity of its elections" and protecting voters from undue influences, confusion, or fraud was indeed compelling, but nevertheless, the laws were constitutionally deficient for several reasons. The hearing and enforcement timeline was long enough that there may not be verdicts rendered until after the election, when it would be too late. There was no procedure to "screen out frivolous complaints" prior to a probable cause hearing, which could lead to gamesmanship by political opponents, who could file complaints shortly before an election, knowing the commission

102. *Id.* at 404 (internal citation omitted).
103. *Id.* at 404.
104. Susan B. Anthony List v. Driehaus, 814 F.3d 466 (6th Cir. 2016).
105. *Id.* at 470.
106. OHIO REV. CODE ANN. § 3517.21(B)(10) (LexisNexis 2024).
107. *See Susan B. Anthony List*, 814 F.3d at 471.
108. *Id.* at 473.

could not resolve the matter before the election.[109] The law was also so broad as to "apply to *all* false statements, including non-material statements," (emphasis added) meaning harmless statements could trigger investigations and penalties.[110] The court ultimately said the law was "both over-inclusive and underinclusive" to advance the interests in promoting fair elections, referencing both *Commonwealth v. Lucas* and *281 Care Committee* in its reasoning. Complaints can cause harm to campaigns through preliminary probable-cause hearings about statements that may not be in violation of the law, while also failing to timely penalize those who violate the law, nor providing remedies for campaigns that are the victims of such "damaging false statements."[111]

Thus, if AI-generated videos, audio, photographs, and other deepfakes are classified as false political speech, laws attempting to curtail or ban them will face daunting constitutional challenges. Political speech is at the core of the First Amendment's purpose and faces the highest level of protection. False speech in general is protected under the Supreme Court's decision in *United States v. Alvarez* in 2012, with states having to establish "legally cognizable harm" and overcome the strict scrutiny standard by demonstrating a compelling government interest and a regulation narrowly tailored to advance that interest.[112] In the context of false political speech, the final appellate decisions in every case decided after *Alvarez* have been consistent in finding flaws in the challenged statutes. Practically, the laws invited gamesmanship and political trickery by allowing anyone to complain to trigger investigations that were implausible to conduct in a thorough and meaningful manner within the compressed timeline of a heated election. And, as a matter of First Amendment scrutiny, the laws failed because of what the *Alvarez* decision identified, and the lower courts also embraced, as the least restrictive remedy for false political speech: counterspeech.[113]

This sets the foundation for review as we examine efforts by states specifically targeting deepfakes and AI-generated false political speech.

C. First Amendment Analysis of Legislation Aiming to Ban or Limit Synthetic Media Used in Political Campaigns

We begin by reviewing the work of commentators and legislators regarding the regulation of deepfakes and AI-generated

109. *Id*. at 474.
110. *Id*. at 475.
111. *Id*.
112. *See* United States v. Alvarez, 567 U.S. 709, 719 (2012).
113. *Id.* at 727.

media, followed by an analysis of common provisions in state laws. These include a variety of issues in their mechanisms aimed at curbing false political speech, including (a) bans on use of deepfakes within a certain timeframe before an election, (b) requiring disclaimers or disclosures identifying deepfakes or manipulated media, (c) carveouts for areas such as satire, parody, news media coverage, and speech made without requisite scienter, (d) injunctive relief, and (e) enforcement challenges.

As machine learning and artificial intelligence technology were advancing in 2019, Robert Chesney and Danielle Citron explored numerous ways in which the emerging problem of deepfakes – "technologies for altering images, video, or audio (or even creating them from scratch)" – may raise legal issues.[114] They noted the challenges of regulation of false speech following the court's decision in *Alvarez*, which "would seem to preclude a sweeping ban on deep fakes," though they also observed that it left "considerable room for carefully tailored prohibitions of certain harmful deepfakes."[115] Among these were civil liability through lawsuits in areas such as the tort of intentional infliction of emotional distress, violation of the right of publicity, or copyright law.[116] To curb potential spread on social media, they urged revising Section 230 of the Communications Decency Act to "allow a limited degree of platform liability relating to deepfakes."[117] But they noted difficulties, both constitutional and practical, in criminalizing deepfakes in the context of political speech, as malign foreign actors and intelligence services – potentially the worst abusers of such material – would likely remain undeterred. "Ultimately," they concluded, "criminal liability is not likely to be a particularly effective tool against deepfakes that pertain to elections."[118]

Nevertheless, legislators began targeting the looming threat of deepfakes. Senator Ben Sasse introduced the Malicious Deep Fake Prohibition Act in 2018, but the bill did not advance in Congress. The bill defined "deepfake" as "an audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual" and made it a felony to distribute deepfakes that "facilitate criminal or tortious conduct under Federal, State, local, or Tribal law."[119] Nina Brown, who authored

114. Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1757 (2019).

115. *Id*. at 1791.

116. *Id*. at 1792–94.

117. *Id*. at 1799.

118. *Id*. at 1804.

119. Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong., 2d Sess. (2018).

a foundational review of early attempts to regulate deepfakes, noted that "all of the conduct prohibited under the draft bill is already prohibited under existing law – it just further criminalizes those wrongs."[120]

In June 2019, Texas passed the first law of its kind specifically targeting deepfake videos. S.B. 751 amended the Texas Election Code to make it a criminal offense for a person, "with intent to injure a candidate or influence the result of an election, (1) creates a deep fake video; and (2) causes the deep fake video to be published or distributed within 30 days of an election." The legislature defined a "deepfake video" as "a video created with artificial intelligence that, with the intent to deceive, appears to depict a real person performing an action that did not occur in reality."[121] Violations are classified as a Class A misdemeanor, punishable by up to a year in jail and up to a $4,000 fine.[122]

California also passed a pair of laws in 2019 targeting deepfakes.[123] AB 602 provided a right for individuals to sue over sexually explicit deepfakes,[124] and AB 730, which updated the state's "Truth in Political Advertising Act," establishing civil remedies, including injunctive relief for candidates if a person were to "distribute, with actual malice, materially deceptive audio or visual media…of the candidate with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate" within sixty days of an election – essentially, an enhancing state libel law.[125] The law specifically exempted "satire and parody" as well as news media coverage, and it allowed such materials to be distributed if they included a "disclosure stating 'This [image/video/audio] has been manipulated,'" with the appropriate term inserted in the blank. The bill was initially scheduled to sunset in 2023, but the legislature extended that date to 2027.[126] A new round of bills was passed in California in 2024, including the Defending Democracy from Deepfake Deception Act, requiring online platforms to label and block "materially deceptive" content, anticipating that in 2024, "disinformation powered by AI will pollute our information ecosystems like never before."[127]

Here, the authors will review some common language and issues in these laws to see how they may comport with the First Amendment standards detailed in the analysis of RQ2.

---

120. Brown, *supra* note 50, at 49.
121. TEX. ELEC. CODE ANN. § 255.004 (West 2024).
122. TEX. PENAL CODE ANN. § 12.21 (West 2024).
123. Brown, *supra* note 50, at 46.
124. A.B. 602 Cal. Assemb., Reg. Sess. (2019), § 4.
125. A.B. 730, 2019-20 Cal. Assemb., Reg. Sess. (2019), § 3.
126. A.B. 972, 2021-22 Cal. Assemb., Reg. Sess. (2022).
127. A.B. 2655, 2023-24 Cal. Assemb., Reg. Sess. (2024), § 3.

1. Electioneering

These are common in state legislative efforts to restrict political speech involving deepfakes or other AI-generated content. Texas, as noted above, imposes a 30-day limit on the distribution of "deepfake videos" before an election,[128] while California has a 60-day restriction.[129] Michigan's laws regulating AI use in elections and Minnesota's ban on deepfakes aimed at injuring a candidate[130] – both passed in 2023 – further limit such speech by prohibiting their use within 90 days of an election.

Such clauses, generally thought of as "electioneering," are similar to the issue addressed by the Supreme Court in *Citizens United v. FEC* in 2010. When the Court overturned the Bipartisan Campaign Finance Reform Act of 2002, it repeatedly emphasized that the Act limited speech within thirty days of a primary and sixty days of a general election; one challenged was that Citizens United wished to air its documentary about Hillary Clinton within thirty days of the 2008 presidential primary via video-on-demand.[131] The majority opinion by Justice Kennedy offered several examples of actions that would have been banned under criminal sanction but that instead should be protected as political speech, including: "The Sierra Club runs an ad, within the crucial phase of sixty days before the general election, that exhorts the public to disapprove of a Congressman who favors logging in national forests." This was among several examples of what Justice Kennedy called "classic censorship," part of what led the majority to strike down almost all of the law's restrictions on political speech by corporations and unions.[132] Justice Stevens, in his dissent, emphasized that all Citizens United wanted was to get around the 30-day limit on its speech, which Stevens viewed as a valid "time, place, or manner restriction" for speech covering "a narrow subset of advocacy messages…made during discrete time periods."[133] While this may have been enough for the dissenters, it clearly was not for the controlling majority.

As detailed above in the *281 Care Committee* and *SBA List* cases, the timing of complaints matters. In those cases, judges detailed the challenge to adjudicate complaints in a timely manner during a heated election. A complaint raised within thirty or even sixty days of an election is unlikely to be meaningfully addressed by courts or commissions, while the harm of filing such complaints

128. TEX. ELEC. CODE ANN. § 255.004 (West 2024).
129. CAL. STAT. ch. 493, § 4 (AB 730) (2019).
130. MINN. STAT. § 609.771(2) (2024).
131. Citizens United v. FEC, 558 U.S. 310, 321 (2010).
132. *Id*. at 337.
133. *Id*. at 419 (Stevens, J., concurring in part and dissenting in part).

– potentially without merit or as an act of gamesmanship – could go unremedied. As such, the electioneering time limits are unlikely to pass First Amendment muster.

2. Disclaimers and Disclosures

While the *Citizens United* decision provides little help for time limits on electioneering bans, the majority took a more favorable view of disclaimer and disclosure requirements. While "disclaimers and disclosure requirements may burden the ability to speak," Kennedy wrote, they do not suppress speech entirely but may still be permissible unless they open contributors to harassment, create fear of reprisal from the government, or otherwise chill political speech.[134]

Disclosures are among the most common requirements in state deepfake and AI laws, and they may be the most constitutionally permissible. California, Idaho, Indiana, Michigan, New Mexico, Utah, Washington, and Wisconsin have enacted laws that include disclosure provisions.

In some instances, the law centers on mandated disclosure. For example, Michigan's law, enacted in 2023, specifically targets AI manipulations of political speech,[135] and requires "paid political advertisements" to include language that the advertisement "was generated in whole or substantially by artificial intelligence," with special details about how the disclosure must be made depending on whether the advertisement is graphic (including photo and video) or audio.[136] A first violation of the law comes with a civil fine of up to $250, with additional infractions facing fines of up to $1,000.[137] Likewise, Wisconsin's new law, which was enacted in March 2024, requires phrases such as "Contains content generated by AI" or "This video content generated by AI" for audio and video content that the statute defines as "Synthetic media."[138] The law comes with a civil penalty of up to $1,000 for violations.[139] In other instances, disclosure is an affirmative defense, such as in Idaho's "Freedom from AI-Rigged (FAIR) Elections Act," enacted in 2024,

134. *Id*. at 366–67.

135. MICH. COMP. LAWS § 169.202(1) (2023). Artificial intelligence is defined as: "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments, and that uses machine and human-based inputs to do all of the following: (a) Perceive real and virtual environments; (b) Abstract such perceptions into models through analysis in an automated manner; (c) Use model inference to formulate options for information or action."

136. MICH. COMP. LAWS § 169.259(1) (2024).

137. MICH. COMP. LAWS § 168.259(2) (2024).

138. WIS. STAT. § 11.130(2m) (2024).

139. *Id*.

which allows individuals accused of using synthetic media to rebut any civil action by including a prominent disclosure stating, "This (video/audio) has been manipulated," as detailed in the statute.[140]

Generally, the Constitution abhors compelled speech, and the Supreme Court has expressed concerns about deterring participation in groups through mandated disclosure of individuals' names, as it did when supporting the NAACP's right under the Fourteenth Amendment not to be forced to disclose its membership to the Alabama government in 1958.[141] But in *Buckley v. Valeo*, decided in 1976, the Supreme Court upheld disclosure requirements in the context of campaign contributions, finding that they directly serve the governmental interest of providing "the electorate with information" about where political contributions are coming from, while also deterring corruption.[142] As long as the requirement is no more restrictive than necessary, provisions in AI and deepfake laws regulating political speech could survive the strict scrutiny analysis. Because the disclosures are about the nature of the content, rather than names of individuals generating them or distributing them, the potential chilling effect on speech would be minimal, further supporting the constitutionality of such provisions.

However, disclosure requirements may be overly burdensome. California's AB 2839, targeting deceptive media in campaign advertisements, requires that visual media include disclosures that are "easily readable by the average viewer and no smaller than the largest font size of other text appearing" in the work, and the disclosure must appear throughout the entirety of the video. The Eastern District of California found this requirement "overly burdensome" in issuing a preliminary injunction against the law, explaining that such requirements "in this case and many other cases would take up an entire screen, which is not reasonable because it certainly 'drowns out' the message the satire or parody video is trying to convey."[143]

While narrower disclaimers and disclosure requirements may be legally valid, they nevertheless may prove practically unenforceable and even "ripe for abuse," as Alexandra Tushman noted when reviewing California's law aimed at deepfakes.[144] No malicious actor would use disclaimers properly anyway, with little fear of enforcement. Additionally, "if a deepfakes creator released the original, unaltered form of a video with the manipulated

140. IDAHO CODE § 67-6628A (2024).
141. NAACP v. Alabama *ex rel*. Patterson, 357 U.S. 449 (1958).
142. Buckley v. Valeo, 424 U.S. 1, 66–67 (1976).
143. Kohls v. Bonta, 752 F. Supp. 3d 1187, 1197 (E.D. Cal. 2024).
144. Tushman, *supra* note 3, at 1420.

disclaimer, and the manipulated version without it, this would inevitably sow confusion among media consumers."[145]

3. Satire, Parody, News Media, and Scienter

Drafters of these state laws were largely cognizant of potential First Amendment challenges facing restrictions on false political speech, as is evident from the savings clauses included in most of the laws. These provisions exempt the laws from being applied to areas traditionally protected by the First Amendment.

For example, Michigan's law has exemptions for "satire and parody,"[146] as do laws in California[147] and New Mexico. As the Supreme Court noted in *Hustler Magazine v. Falwell* (1988), when the Court held that the First Amendment barred an intentional infliction of emotional distress tort claim by a high-profile public figure for an ad parody making fun of him, explaining that "graphic depictions and satirical cartoons have played a prominent role in public and political debate."[148] While satire and parody may be outrageous, they receive heightened protection when the target is public figures, such as the famous evangelist Jerry Falwell. Laws without such a savings clause for parody and satire would face facial overbreadth challenges.

Likewise, most state laws on deepfakes and AI-generated false speech include exceptions for news media and broadcasters that air paid political advertising.[149] These efforts to protect what states recognize as bona fide news efforts to cover –and potentially debunk –false political speech generated using AI tools are necessary for preserving them from First Amendment challenges.

The Georgia bill exempts not only "satire, parody, works of artistic expression, or works of journalism by bona fide news organizations," but also "activities protected by the First Amendment to the United States Constitution."[150] This circular provision essentially acknowledges that the First Amendment already protects much, if not all, of the speech covered by the act. This provision may have the effect of invalidating the entirety of the law, except possibly the mandatory disclosure provision, as nearly all the remaining portions would likely receive heavy protection under the First Amendment.

---

145. *Id*.

146. MICH. COMP. LAWS § 169.259(4) (2024).

147. CAL. ELEC. CODE § 20010(d)(5) (West 2024).

148. Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 54 (1988).

149. *See, e.g.,* CAL. ELEC. CODE § 20010(d)(4) (West 2024); N.M. STAT. ANN. 1-19-26.4(G); WIS. STAT. § 11.1303(2m)(g) (2024).

150. H.B. 986, 157th Gen. Assemb., 2d Reg. Sess. § 3 (Ga. 2023).

Similarly, many states include heightened scienter requirements to address the Supreme Court's decisions in *New York Times v. Sullivan* and its progeny, which require public officials and public figures to prove actual malice —knowledge of falsity or reckless disregard for the truth— by convincing evidence to recover tort damages based on false political speech. California explicitly requires a showing of "actual malice" for the distribution of "materially deceptive audio or visual media," for example.[151] Laws that do not require such a showing, or likewise that do not require that plaintiffs establish some sort of harm, would face facial challenges under *Sullivan*.[152]

4. Injunctive Relief

While remedies vary across the states, ranging from civil fines to criminal penalties, most state laws targeting deepfakes and AI-generated false political speech provide injunctive relief as a remedy. In the First Amendment context, injunctions and restraining orders are classified as prior restraints on publication, and the Supreme Court has been consistently hostile to them in the context of political speech for generations.[153] As the Court established in *Bantam Books, Inc. v. Sullivan* (1963), "Any system of prior restraints of expression comes to this Court bearing a heavy presumption against its constitutional validity."[154] Indeed, the first court to review a political deepfake law found the prior restraint to be a concern. The Eastern District of California noted that "the First Amendment was designed to protect citizens against prior restraints and encroachments of speech by State governments themselves."[155]

For laws in which the only requirement is the mandatory disclaimer or disclosure that campaign speech has been generated by AI, such as New Mexico's act,[156] injunctive relief might be a more

151. CAL. ELEC. CODE 20010(a) (West, 2024).

152. *See* Ex parte Stafford, 667 S.W.3d 517, 532 (Tex. Ct. App., 2023) (holding the Texas "True Source of Communication" law violated the First Amendment even with a scienter requirement of actual malice, which "do little, if anything, to narrow the extensive reach of this statutory element before an individual is charged…The assessment of whether a communication emanates from its true source is left to prosecutorial discretion. Thus, the mens rea requirements do not mitigate or eliminate the risk of chilling protected speech or guard against the danger of arbitrary and discriminatory enforcement.").

153. Near v. Minnesota, 283 U.S. 697 (1931); N.Y. Times v. United States, 403 U.S. 713 (1971).

154. Bantam Books, Inc. v. Sullivan, 372 U.S. 58, 70 (1963).

155. Kohls v. Bonta, 752 F. Supp. 3d 1187, 1199 (E.D. Cal. 2024).

156. N.M. STAT. ANN. § 1-19-26.4(F) (2024); *see also* N.M. STAT. ANN. § 1-19-34.6 (2024) (permitting injunctive relief to enforce campaign regulations).

plausible remedy, as it would only compel the creator or distributor of the content to include the required disclosure language.

But in most states, injunctive relief appears to be offered as a broad remedy for whoever files a complaint. This relief grants plaintiffs the ability to seek a court order to take down and stop the distribution of content violating the deepfake or AI-generated content bans. For example, Washington's law allows a "candidate whose appearance, action, or speech is altered through the use of synthetic media in an electioneering communication" to "seek injunctive relief or other equitable relief prohibiting the publication of such synthetic media."[157] Such a ban, barring other heightened proof requirements for movants seeking such an injunction, does not seem to comport with First Amendment doctrine. False political speech has First Amendment protection, and mere falsity in the absence of significant harm or danger to individuals or the state is not enough to clear the bar for prior restraints. The Supreme Court, of course, allowed the publication of classified material regarding the Vietnam War in the Pentagon Papers case on grounds that they did not imminently jeopardize national security.[158] Political campaign speech, even when false or defamatory, has stronger First Amendment protection.

It is plausible that a restraining order or injunction might be valid upon a showing of defamation and harm to a public figure. Here, the plaintiff must meet the actual malice standards and heightened burden of proof required in *Sullivan*, through what have become known as "anti-libel injunctions."[159] Courts have permitted such injunctions after a final judgment determines that the speech targeted was indeed defamatory. Although such injunctions are limited to past speech, not future conduct, limiting relief to takedowns only and not recirculation of the defamatory content.[160]

5. Enforcement Issues

A final challenge to address in state laws aimed at deepfakes and AI-generated false political speech is enforcement. While Texas and California have had laws in place specifically targeting political deepfakes since 2019, the authors were unable to find evidence of any prosecutions or lawsuits under those laws reported in the four years since. As detailed in the analysis of RQ2, prosecution of false political speech is a treacherous area, ripe for potential abuse and

157. WASH. REV. CODE § 42.62.2020(2) (2024).
158. N. Y. Times v. United States, 403 U.S. 713 (1971) (Black, J., concurring).
159. *See* Eugene Volokh, *Anti-Libel Injunctions*, 167 U. PA. L. REV. 73 (2019).
160. *Id.*; *see also* Kinney v. Barnes, 443 S.W.3d 87 (Tex. 2014).

selective prosecution. And courts reviewing similar provisions regarding false political speech, on their face or as applied, have found numerous flaws that rendered them inapplicable under the First Amendment.

One flaw in many of the laws is that nearly anyone can make a complaint about a deepfake or an AI-generated photo or video. This opens the floodgates to abuse through frivolous filings and political gamesmanship. As Chesney and Citron noted, minority or unpopular viewpoints might be targeted by such laws, resulting in "politicized enforcement" and "inhibit[ed] engagement in political discourse" through the resulting chilling effects.[161] For example, in California, "any registered voter may seek a temporary restraining order and an injunction prohibiting the publication…of any campaign material in violation of this section."[162] The Michigan law may be enforced by the Attorney General, a candidate claiming injury, a depicted individual, or "any organization that represents the interests of voters likely to be deceived" by the manipulated content.[163] It is more likely that laws that limit the pursuit of remedies to candidates, such as Idaho,[164] or the state elections commission, such as Wisconsin,[165] would fare better under First Amendment review.

When examining the topic of deepfakes in a First Amendment context, courts have so far been skeptical that the purported harms are plausible. This has generally been in the context of cases where courts have struck down bans on audio/video recording or broadcasting. The Ninth Circuit Court of Appeals in 2023 struck down an Oregon law prohibiting recording audio or video without consent, dismissing an argument by the dissent that the law would be helpful in preventing deepfakes using such audio/video.[166] Similarly, in 2022, the Maryland District Court struck down the "Broadcast Ban" rule that barred distribution of audio/video taken in criminal courtrooms, even though defenders of the ban worry that the audio/video could be used in deepfakes.[167]

The one court to review a false political speech law that contains a deepfake ban was similarly skeptical. In *Ex parte Stafford*,[168] a Texas state appeals court ruled in 2023 that the Texas "True Source of Communication" law —which includes the "deep fake video" ban passed in 2019—was unconstitutionally

161. Chesney & Citron, *supra* note 114, at 1789–90.
162. CAL. ELEC. CODE § 2100.10(c)(1).
163. MICH. COMP. LAWS SERV. § 168.932f(4) (LexisNexis 2024).
164. *See* IDAHO CODE § 67-6628A(3) (2024).
165. *See* WIS. STAT. § 11.1303(2m)(d).
166. Project Veritas v. Schmidt, 72 F.4th 1043 (Ninth Cir. 2023).
167. Soderberg v. Carrion, 645 F. Supp. 3d 460 (D. Md. 2022).
168. *Ex parte* Stafford, 667 S.W.3d 517 (Tex. App. 2023).

overbroad, although the challenge was directed at a different provision. John Morgan Stafford was indicted under the law for "sending text messages with the appearance of coming from a Republican or conservative campaign."[169] This violated section (b) of the law, which forbids "knowingly represent[ing] in a campaign communication that the communication emanates from a source other than its true source."[170] Applying strict scrutiny, the court found that Texas had a compelling interest in "promoting honest discourse and preventing misinformation in the political arena."[171] But the court found the statute was not narrowly tailored to achieve that interest because the state offered no empirical evidence beyond "common sense" to support the necessity of the statute, which is inadequate to meet First Amendment burdens.[172] And the court referenced *Alvarez* for the "generally accepted proposition that counter speech may provide a less restrictive means of advancing the state's interest," as one of several less restrictive means available to the state.[173] Possible less restrictive means included the misrepresentation statute in the Election Code, which makes it a crime for a person to misrepresent themselves as a candidate or as an agent of a candidate.[174] Ultimately, the court ordered the dismissal of charges against Stafford.

Thus, common provisions in state laws regarding deepfakes and AI-generated speech regarding elections face significant First Amendment hurdles. While mandatory disclosures and disclaimers seem most likely to withstand scrutiny by courts, other provisions, such as injunctive relief, are less likely, and laws without heightened protection for satire, parody, news media, and scienter could be constitutionally deficient.

CONCLUSION

As we were finalizing this article, a headline in the *Washington Post* caught our attention: "Deepfake Kari Lake Video Shows Coming Chaos of AI in Elections." But in reality, there was no chaos. The video was a ploy by a journalist for the *Arizona Agenda* to show the potential harm of AI-generated videos. It was inspired, the journalist said, by a similar set of videos created by a state elections official for training staff. The synthetic Kari Lake, a candidate for the U.S. Senate, was generated to say: "Subscribe to the Arizona Agenda for hard-hitting real news… And a preview of

169. *Id.* at 521.
170. TEX. ELEC. CODE § 255.004(b) (West 2024).
171. *Stafford*, 667 S.W.3d at 525.
172. *Id.* at 528.
173. *Id.*
174. *Id.* at 527 (citing TEX. ELEC. CODE § 255.005 (West 2024)).

the terrifying artificial intelligence coming your way in the next election, like this video, which is an AI deepfake the Arizona Agenda made to show you just how good this technology is getting."[175]

There appears to be no shortage of fears of a voting public duped by deepfakes and other AI-manipulated photos, videos, and audio. One of the earliest political deepfake videos to get widespread attention was generated by Jordan Peele in 2018, depicting Barack Obama insulting Donald Trump using vulgar language as a warning to "raise public awareness about deepfakes."[176] Likewise, the synthetic audio mimicking the voices of legislators opposing the AI-generated campaign speech bill in Georgia in 2024 was done as a cautionary tale to encourage voting for the bill restricting such materials, rather than an authentic attempt to manipulate citizens about issues in an election.

The highest-profile efforts to manipulate video, audio, or photo to portray politicians in a false manner–consider the deceptively edited Nancy Pelosi video in 2019, or the fake audio of Joe Biden circulating before the 2024 New Hampshire primary, the AI-generated photos of Donald Trump among fictitious Black supporters, or the parody Kamala Harris video–not only have been caught but also quickly criticized by political opponents and news media. Perhaps the most successful AI-generated photo hoax yet, of Pope Francis wearing a puffy, white, expensive Balenciaga coat that circulated in March 2023, may have "fooled us all," but was debunked as a fake and covered extensively across popular media within days.[177]

The debunking of these deceptions by citizens, campaigns, and news media provides the counterspeech that U.S. courts have identified time and again as the most effective and least speech-restrictive remedy for false political speech. In light of the Supreme Court's decision in *U.S. v. Alvarez* (2011) and the way in which lower courts, including the Sixth and Eighth Circuit Courts of Appeals have applied *Alvarez* to political false speech regulations, it is hard to overcome counterspeech as an easier and more accessible option than bans, criminal penalties, or injunctions. Narrower existing laws regarding misrepresentation of oneself as a public official, safely allowable under *Alvarez*, would still be in place

175. Reis Thebault, *Deepfake Kari Lake video shows coming chaos of AI in elections*, WASH. POST (Mar. 24, 2024), https://www.washingtonpost.com/politics/2024/03/24/kari-lake-deepfake/ [https://perma.cc/H8CN-LPDZ].

176. Brown, *supra* note 50, at 5.

177. Ashley Fetters Maloy & Anne Branigin, *An AI-generated 'Balenciaga pope' fooled us all. How much does it matter?*, WASH. POST (Mar. 27, 2023), https://www.washingtonpost.com/lifestyle/2023/03/27/pope-francis-coat-puffy-white-ai-fake/ [https://perma.cc/KQ6S-NQ4A].

to handle audio manipulations and impersonations such as the deceptive Biden phone call in New Hampshire.

While it is possible that the Supreme Court could be coaxed into reconsidering *Alvarez*–only three of the six justices who found that false speech was protected and required striking down the Stolen Valor Act remain on the Court–the Court has been reluctant to create new categories of unprotected speech.[178] Supporters of deepfake and AI generated political false speech laws would need to show counter speech and other remedies—such as disclosures, disclaimers, defamation and right of publicity—were inadequate. They would have to demonstrate that these limits leave the harm caused by synthetic media in the political context unaddressed. Further, they would have to establish that the bans or limits would directly advance the interest in free and fair elections targeted by such laws.[179]

A better path to managing deepfakes and AI-generated false political speech, practicality and legally, may be rooted in technology and the free market. Social media platforms have tried to fight synthetic media disinformation using content moderation strategies, some of those strategies have included AI. However, there are other technical solutions possible, as imagined by AI companies. In 2023, Google announced its SynthID digital watermarking tool that would encode the knowledge of an AI image's provenance at the point of creation.[180] However, one problem with this is that the encoding process was only available to those using Google's image generation tool because it wasn't interoperable across platforms.[181] Still, Google was one of seven

178. *See* United States v. Stevens, 559 U.S. 460, 472 (2010) ("Our decisions in *Ferber* and other cases cannot be taken as establishing a freewheeling authority to declare new categories of speech outside the scope of the First Amendment. Maybe there are some categories of speech that have been historically unprotected, but have not yet been specifically identified or discussed as such in our case law.").

179. *See* Matthew Bodi, *Note, The First Amendment Implications of Regulating Political Deepfakes*, 47 RUTGERS COMPUT. & TECH. L.J. 143, 169 (2023) ("Any counterspeech will be incredibly difficult because of the future difficulty, or some would even argue the futility, in detecting deepfakes. In addition, the widely varying channels of dissemination of deepfakes make it difficult for any counterspeech to effectively compete against it.").

180. David Pierce, *Google made a watermark for AI images that you can't edit out*, VERGE (Aug. 29, 2023, 6:00 AM MDT), https://www.theverge.com/2023/8/29/23849107/synthid-google-deepmind-ai-image-detector [https://perma.cc/6LGY-HG9Z].

181. Garrit De Vynck, *AI images are getting harder to spot. Google thinks it has a solution*, WASH. POST (Aug. 29, 2023), https://www.washingtonpost.com/technology/2023/08/29/google-wants-watermark-ai-generated-images-stop-deepfakes/ [https://perma.cc/8MDU-RJBD].

major U.S. AI companies that announced they would voluntarily work to watermark images produced by its tools.[182]

The hope with watermarking is that by making it part of the code, it would be easier for posting platforms to scan the image and label AI content appropriately at the point of publication rather than long after the fact. This is precisely what Facebook's parent company, Meta, announced it would do in 2024 when it said it would begin to label AI content from Google and OpenAI's DALL-E platforms.[183] What AI watermarking lacks is a common standard similar to how Exif data is used to embed data such as camera type or location into the code of a digital photo. Without a common, open-source system for AI companies, the publishing platforms will have to keep adapting their labeling solutions to several competing standards that may arise to address the problem, making the solution less powerful. One other potential pitfall to watermarking is if those intending to use generated images for misinformation figure out a workaround to fool publishing platforms (such as screenshotting an image, which is what people do to strip an image of Exif data for example), ensuing images that are AI fakes would potentially be unlabeled and make misinformation worse if users learn to trust the labeling system and have lower skepticism of unlabeled AI fakes.

A bill introduced to Congress in 2024 considered the question of watermarking. HR 6466, known as "The AI Labeling Act," would require any system that generates images, video, audio, or multimedia to include a disclosure about the digital artifact's AI provenance. The act would also require the output to include metadata that notes it is AI content, to identify the platform used to make the digital artifact, and to timestamp its creation date. In perhaps a nod to the reality that the legislation is getting ahead of an AI company's technical know-how, the act would require that these disclosures "shall, to the extent technically feasible, be permanent or unable to be easily removed by subsequent users."[184]

So far, in light of the First Amendment doctrine that provides a strong protection for false speech and political speech, labeling and disclosure requirements are likely the most feasible way to

182. Bill Rosenblatt, *Google and OpenAI Plan Technology to Track AI-Generated Content*, FORBES (July 22, 2023, 12:06 EDT), https://www.forbes.com/sites/billrosenblatt/2023/07/22/google-and-openai-plan-technology-to-track-ai-generated-content/?sh=3ce26836131b [https://perma.cc/DQY2-L2R9].

183. Benj Edwards, *Meta will label AI-generated content from OpenAI and Google on Facebook, Instagram*, ARS TECHNICA (Feb. 6, 2024, 11:04 AM), https://arstechnica.com/information-technology/2024/02/meta-will-label-ai-generated-content-from-openai-and-google-on-facebook-instagram/ [https://perma.cc/DQY2-L2R9].

184. AI Labeling Act of 2023, H.R. 6466, 118th Cong. (2023).

manage the potential onslaught of deepfake and deceptive AI-generated content that policymakers anticipate. While preventing disinformation and misinformation from disrupting political campaigns is certainly valuable, giving power to the state to adjudicate what is true, what is false, and who should be punished for creating and circulating such content may be an even greater danger. As the Court noted in *Alvarez*, "Our constitutional tradition stands against the idea that we need Oceania's Ministry of Truth."[185]

185. United States v. Alvarez, 567 U.S. 709, 723 (2012).