

The First AI-Saturated Election: Disclosure, Detection, and the Regulatory Limits of Generative Political Persuasion

Feilian Huang*

Yujia Liu†

Abstract

The campaign cycle preceding the November 3, 2026 United States midterm elections is the first in which generative artificial intelligence has functioned as a routine instrument of political persuasion: party committees have aired synthetic videos of opposing candidates, manipulated campaign mailers have circulated at industrial scale, and foreign actors have seeded AI-generated disinformation through fabricated broadcast footage. Legislatures responded with unprecedented speed: as of September 2026, thirty-three states have enacted statutes regulating AI-generated election content, most built on mandatory disclosure, a minority on outright prohibition. Yet the regulatory edifice is already cracking. By September 2026, federal courts had permanently enjoined California’s distributor-liability statute and its platform-takedown mandate, struck down Hawaii’s criminal deepfake law, and preliminarily blocked Montana’s prohibition regime on viewpoint-discrimination grounds. This paper argues that state regulation of generative political persuasion faces two independent constraints that together make speaker- and content-focused regulation unworkable. First, constitutional: the First Amendment doctrines of strict scrutiny for content-based political speech, the prohibition on viewpoint discrimination, and the compelled-speech doctrine place most prohibition statutes, and some disclosure regimes, in serious jeopardy. Second, technical: the detection and provenance mechanisms on which disclosure duties depend, including text watermarking, C2PA metadata, and classifier-based detectors, are fragile against motivated adversaries and unenforceable at electoral scale. Comparing the American experience with the European Union’s AI Act Article 50 transparency regime, which allocates provider-marking and deployer-disclosure duties differently and entered application in August 2026, we propose a different regulatory path: shift emphasis from regulating speakers and content toward preserving provenance signals and imposing duties on distributors and intermediaries. The framework offers a post-midterm legislative agenda grounded in what the technology can actually support.

1 Introduction

Some time in the spring of 2026, the National Republican Senatorial Committee began airing a television advertisement in which Democratic Texas Senate candidate James Talarico appeared to say that “radicalized white men are the greatest domestic terrorist threat in our country.” The words were never spoken. The video was generated by artificial intelligence, and the only indication, a two-word disclaimer in small type in the lower corner of the frame, was easy to miss ([Reuters, 2026a](#)). By September 2026, a Texas Republican Senate nominee had posted an AI-generated video of the same candidate writing “let there be six genders” in a Bible with crayons, appended with a donation QR code ([AFP, 2026](#)); a Michigan Senate hopeful was circulating as a Hulk-like figure rescuing citizens from a collapsing building ([Mint, 2026](#)); and a Montana political action committee

*Independent Researcher. ORCID: [0009-0006-2214-0233](#).

†Independent Researcher.

was mailing voters AI-generated images of candidates holding pride flags and wearing pronoun pins, images that would, within days, become the subject of a federal court order (Reuters, 2026b). United States intelligence assessments, reported in mid-September 2026, described a Russian AI-driven influence campaign fabricating celebrity videos that mimicked CNN, BBC, and *New York Times* broadcasts about Senate races, already being seeded on social media (DISA, 2026).

The campaign cycle preceding the November 3, 2026 midterm elections is thus the first in which generative AI has operated not as a curiosity or a scandal but as ordinary campaign infrastructure. The scale is difficult to overstate: one cybersecurity firm estimated that deepfakes shared on social media grew from roughly 500,000 in 2023 to 8 million in 2025, and a September 2026 poll found that 85 percent of Americans expect AI-created political content to spread misleading information about the midterms (Mint, 2026). There is no federal regulation constraining the use of AI in American political messaging; in its place stands a patchwork of largely untested state laws, in the words of one Reuters dispatch, “a media landscape with few guardrails” (Reuters, 2026a).

State legislatures have not been idle. As of September 3, 2026, thirty-three states had enacted laws regulating deepfakes used in political communications, up from twenty-eight at the start of the year, according to Ballotpedia’s annual deepfake-legislation report (Ballotpedia, 2026). The designs vary: most states require disclosure that content was AI-generated; a minority prohibit deceptive synthetic media outright during election periods (Washington Examiner, 2026). But the word “enacted” must be read with care, because enactment has not meant enforcement. In August 2025, a federal court permanently enjoined California’s private-right-of-action statute for deceptive election media on First Amendment grounds, and in a separate order enjoined the state’s platform-takedown mandate as preempted by Section 230 (Kohls v. Bonta, 2025; Justia, 2025). In January 2026, a federal judge in Hawaii permanently struck down that state’s criminal deepfake law as content- and speaker-based discrimination that failed strict scrutiny (Bloomberg Law, 2026). And on September 17, 2026, a Montana federal court preliminarily blocked enforcement of that state’s prohibition statute, finding it engaged in viewpoint discrimination by punishing synthetic media that injures a candidate while permitting identical media that elevates one (Reuters, 2026b).

This paper makes three claims. First, the 2026 injunction record, read together, establishes a doctrinal pattern: prohibition-style regulation of synthetic political speech faces severe constitutional headwinds, and even disclosure mandates sit on thinner First Amendment ice than legislators assume, because mandatory AI labels are compelled speech (McIntyre, 1995). Second, the technical premise of disclosure regulation, that AI-generated content can be reliably detected or its provenance reliably established, does not hold at electoral scale against motivated adversaries: text watermarks are removable by automated paraphrase and learnable by query access, C2PA metadata is routinely stripped by platform recompression, and classifier-based detectors combine high false-positive rates with trivial evasion (Sadasivan et al., 2024; Jovanović et al., 2024; Liang et al., 2023; Krishna et al., 2023). Third, a comparison with the European Union’s AI Act, whose Article 50 transparency obligations became applicable on August 2, 2026, reveals a design error in the American approach: the United States has loaded duties onto speakers and distributors of *content*, while the European model more carefully distinguishes the duties of model *providers* (machine-readable marking) from those of *deployers* (disclosure), and even that model depends on technical mechanisms of limited robustness (European Commission, 2026; CSA, 2026a).

The constructive argument follows from the two constraints. If speakers cannot constitutionally be made to label or suppress synthetic political speech, and if synthetic speech cannot technically be detected at scale, then regulation aimed at speakers and content is a dead end. The viable path runs through distribution: legal duties on platforms and intermediaries to preserve provenance signals, to label content whose provenance is verifiably synthetic, and to maintain auditable records of large-scale synthetic political advertising, coupled with distributor-side liability that does not

depend on adjudicating the truth of any particular message. This is not a counsel of despair. It is an attempt to ground election-integrity regulation in what the technology can actually support, so that the post-midterm legislative agenda is built on mechanisms that survive both constitutional review and adversarial pressure.

The paper proceeds as follows. Section 2 surveys the 2026 state regulatory landscape and the injunction record. Section 3 analyzes the First Amendment constraints. Section 4 examines the technical limits of detection and provenance. Section 5 compares the EU AI Act’s Article 50 regime. Section 6 proposes the distributor-duties framework. Section 7 situates the argument in the existing literature, and Section 8 concludes.

A note on method and scope. This paper is a work of legal-technical synthesis, not an empirical study: its evidence base is the published case law through September 2026, the peer-reviewed literature on watermarking and detection, and official EU regulatory sources, as documented in the citations throughout. Its geographic scope is the United States, with the European comparison included for its allocation logic rather than as a claim about European outcomes, which remain untested. Federal proposals, including the Protect Elections from Deceptive AI Act discussed by [Loewenstein \(2024\)](#), enter the analysis only as they illuminate the state-law debate. The paper takes no position on whether any particular 2026 campaign advertisement should have been lawful; its concern is whether the regulatory instruments aimed at such advertisements can survive the constitutional and technical constraints they face.

2 The 2026 Regulatory Landscape

2.1 Thirty-three states, two designs

The state legislative response to synthetic political media has been the fastest-moving front in American election law. Ballotpedia’s third annual report on deepfake legislation, announced August 5, 2026, found that forty-eight states had enacted laws addressing sexually explicit deepfakes and, of direct relevance here, that thirty-three states had enacted laws regulating deepfakes used in political communications, up from twenty-eight at the beginning of the year ([Ballotpedia, 2026](#)). (An older figure of thirty-one, drawn from a National Conference of State Legislatures count in late June, is now superseded.) In more than half of those states, the law applies to material distributed within a specified number of days before an election; most include a disclosure exception, while three states’ prohibition-style laws contain no disclosure safe harbor at all ([Ballotpedia, 2026](#)).

The statutes divide into two architectural families. The majority model is *disclosure*: campaign communications containing AI-generated content must carry a disclaimer or label identifying the synthetic origin. Arizona, Florida, New York, Utah (SB 131), Colorado (HB 24-1147), Oregon (SB 1571), and Wisconsin (Act 123) all follow this template, as does California’s surviving AB 2355 ([Washington Examiner, 2026](#); [Recording Law, 2026](#)). A minority of states chose *prohibition*: Texas, Minnesota, and Maryland bar the dissemination of deceptive synthetic election media during defined election periods, with Texas’s ban applying in the thirty days before an election, Minnesota’s from the start of absentee voting or ninety days before a nominating convention, and Maryland’s prohibition framed indefinitely ([Washington Examiner, 2026](#)). Montana’s SB 25 (2025) prohibits knowingly sponsoring a “deepfake” of a candidate or political party within sixty days before voting begins, with a disclosure escape hatch and an exception for satire ([Reuters, 2026b](#)). Testifying before Nevada lawmakers in 2026, NCSL staff summarized the national picture as roughly two dozen enacted states, most disclosure-based with a few prohibition-based, and cautioned that laws should target demonstrable harms such as voter intimidation or coercion rather than “regulating based on speculative effects to electoral prospects,” warning that many statutes would face First Amendment

tests ([Citizen Portal, 2026](#)). That warning proved prescient.

Two statutes deserve individual attention because they anchor the doctrinal analysis. Texas’s 2019 law, the first election-deepfake statute in the nation (Tex. Election Code § 255.004), prohibits creating or publishing a deepfake video within thirty days of an election with intent to injure a candidate; a subsection was later held unconstitutionally overbroad, though the core prohibition was upheld ([Recording Law, 2026](#)). Minnesota’s H.F. 1370 (2023), codified at Minn. Stat. § 609.771, criminalizes disseminating a deepfake knowingly or recklessly within ninety days before an election to influence it, excluding satire and parody; it has survived early challenge only on standing grounds, with X Corp.’s Section 230 claim dismissed without prejudice for lack of Article III standing in December 2025, and Minnesota’s attorney general was still seeking dismissal of X’s amended challenge as of September 2026 ([Bloomberg Law, 2025](#); [LegalNewsFeed, 2026](#)). The claim that Minnesota’s law “survived challenge” should therefore be stated precisely: it survived a standing-based motion, and its merits remain untested.

At the federal level, the vacuum is nearly complete. Congress has enacted no statute regulating AI-generated election content, and the only federal action to date is the Federal Communications Commission’s February 2024 declaratory ruling banning AI-generated voices in robocalls under the Telephone Consumer Protection Act, issued after a synthetic robocall impersonating President Biden targeted New Hampshire primary voters in January 2024 ([FCC, 2024](#)). That episode remains the template regulators fear, but no reputable reporting has documented a synthetic robocall in the 2026 cycle itself; the paper treats robocalls as a risk scenario grounded in the 2024 precedent, not as an observed 2026 phenomenon.

2.2 Synthetic persuasion in the 2026 cycle

The statutes surveyed above were written against a campaign environment that, by September 2026, had made synthetic media unremarkable. In March 2026, Reuters reported that the National Republican Senatorial Committee was airing an AI-generated advertisement depicting Democratic Texas Senate candidate James Talarico appearing to declare that “radicalized white men are the greatest domestic terrorist threat in our country,” with the words “AI generated” rendered in easy-to-miss type in the lower corner of the frame ([Reuters, 2026a](#)). The dispatch’s framing has been widely quoted since: “The realistic video is among a vanguard of ‘deepfake’ advertisements that some campaigns are already deploying ahead of November’s midterm elections,” introduced into “a media landscape with few guardrails” where “there is no federal regulation constraining the use of AI in political messaging, leaving only a patchwork of largely untested state laws” ([Reuters, 2026a](#)).

By June 2026, the genre had diversified beyond attack ads into synthetic spectacle: the *Wall Street Journal* documented Talarico depicted singing “My Favorite Things” in praise of transgender children, Michigan Senate hopeful Mike Rogers rendered as a Hulk-like figure rescuing citizens from a collapsing building, and Los Angeles mayoral candidate Spencer Pratt as Batman saving the city from an incumbent cast as the Joker ([Mint, 2026](#)). In September 2026, Texas Republican Senate nominee Ken Paxton posted an AI-generated video of Talarico writing “let there be six genders” in a Bible with crayons, appended with a donation QR code; Talarico told NBC News the advertisements “don’t add anything to our civic discourse” ([AFP, 2026](#)). The Montana mailers at the center of the *Knudsen* litigation, AI-generated images of candidates with pride flags and pronoun pins, showed the same techniques migrating to direct mail ([Reuters, 2026b](#)).

Two further data points fix the scale. A cybersecurity firm’s estimate put deepfakes shared on social media at roughly 500,000 in 2023 and 8 million in 2025, and a September 2026 NPR/PBS/Marist poll found 85 percent of Americans expecting AI-created political content to spread misleading

information about the midterms (Mint, 2026). And the foreign dimension is no longer hypothetical: around September 18, 2026, U.S. intelligence assessments reported through the *New York Times* described a Russian AI-driven influence campaign fabricating celebrity videos that mimicked CNN, BBC, and *Times* broadcasts about Senate races, already being seeded on social media (DISA, 2026). The regulatory question is thus not whether synthetic persuasion will matter to the 2026 midterms, but whether any legal instrument is capable of governing it.

2.3 The injunction record: California, Hawaii, Montana

By September 2026, four statutes in three states had been blocked by federal courts, and the pattern of the rulings matters more than the count.

California. Two 2024 statutes fell in the consolidated *Kohls v. Bonta* docket in the Eastern District of California. AB 2839, which created a private right of action against distributors of deceptive election media, was preliminarily enjoined on October 2, 2024 and permanently enjoined on August 29, 2025; the court held that the statute likely violated the First Amendment on its face as a content-based restriction that was not narrowly tailored, and that it impermissibly compelled speech (Kohls v. Bonta, 2025). Judge Mendez’s preliminary-injunction order supplied the passage that has since become the canonical judicial rebuke of the genre: “Most of A.B. 2839 acts as 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” (Vermont Legislature, 2026). The plaintiffs were Christopher Kohls (“Mr Reagan”), the satirical outlet The Babylon Bee, and commentator Kelly Chang Rickert, a lineup that underscored the statute’s burden on parody and humor (Justia, 2024). AB 2655, the Defending Democracy from Deceptive Deepfake Ads Act, which imposed reporting and takedown duties on large platforms, was permanently enjoined on August 20, 2025, but on a different ground: the court held it preempted by 47 U.S.C. § 230(c)(1), (c)(2)(B), and (e)(3) as applied to interactive computer service providers, and did not reach the First Amendment claims (Justia, 2025). The distinction is legally significant: AB 2839 fell on speech grounds, AB 2655 on federal preemption of state platform mandates, and commentary that merges the two holdings misstates the record. California’s AB 2355, a disclosure regime for political advertisements, remains in force (Washington Examiner, 2026).

Hawaii. Act 191 (2024, SB 2687) criminalized the reckless distribution of “materially deceptive media” involving candidates during the February-to-November election season, with a disclaimer safe harbor. On January 30, 2026, Judge Shanlyn A.S. Park of the District of Hawaii permanently enjoined the statute on First and Fourteenth Amendment grounds (Bloomberg Law, 2026). The court held that Act 191 discriminated on the basis of both content and speaker, failed strict scrutiny despite the state’s compelling interest in electoral integrity, was unconstitutionally vague, and compelled speech, in that the mandated disclaimer “would distort the meaning and effect of satirical speech.” As reported by Bloomberg Law, the court deemed the statute “presumptively invalid” because it “discriminates based on content and speaker, and, in doing so, restricts constitutionally protected political speech” (Bloomberg Law, 2026). The plaintiffs were again The Babylon Bee, joined by Hawaii resident Dawn O’Brien and represented by Alliance Defending Freedom counsel (ElectionDesk, 2026). One analysis of the decision treats it as a cautionary template for other states, arguing that its reasoning generalizes well beyond Hawaii’s borders (AEI, 2026).

Montana. The most recent and doctrinally sharpest ruling arrived on September 17, 2026. In *Accountability in State Government v. Knudsen*, Judge Susan Watters of the District of Montana preliminarily enjoined enforcement of SB 25 against the plaintiffs, a former state legislator and his political action committee, whose AI-generated campaign mailers depicted candidates holding pride flags and wearing pronoun pins (Reuters, 2026b). The court held the plaintiffs likely to succeed

on their First Amendment claim because the statute constitutes *viewpoint-based* discrimination: it targets synthetic media created “with the intent to injure the reputation” of a candidate or party while permitting identical deceptive media deployed to bolster a candidate. In the court’s words: “Accordingly, because the Act treats identically deceptive media differently based solely on whether its message is weaponized to damage or deployed to elevate a candidate, the Court finds that Plaintiffs are likely to succeed in establishing the Act constitutes viewpoint-based discrimination” (Reuters, 2026b). Judge Watters additionally flagged the mandatory AI disclosure as compelled speech, noting that creators must affix a government-prescribed message rather than merely identify the payer (Reason, 2026). Two caveats attach. First, the injunction is plaintiff-specific, not statewide: Judge Watters expressly declined to issue universal relief, citing the Supreme Court’s 2025 decision in *Trump v. CASA* limiting nationwide injunctions (Reuters, 2026b; *Trump v. CASA*, 2025). Second, it is a preliminary injunction, decided on likelihood of success; but the viewpoint-discrimination analysis, if sustained, is the kind of defect that narrow tailoring cannot cure.

Taken together, the three episodes trace an arc from content-based overbreadth (California), through speaker- and content-discrimination plus compelled speech (Hawaii), to viewpoint discrimination (Montana). Each successive ruling has found a distinct constitutional infirmity in the same legislative project. That arc is the subject of the next section.

3 First Amendment Constraints

The constitutional analysis begins from a proposition no court in the 2026 cycle has questioned: political speech about candidates for office sits at the core of the First Amendment, and content-based restrictions on it are subject to strict scrutiny, requiring that the law be narrowly tailored to serve a compelling governmental interest (Reed, 2015). Electoral integrity is routinely accepted as a compelling interest; the statutes keep failing on the second prong. The question is why the tailoring inquiry proves so unforgiving here, and the answer lies in four doctrines that interact to close off nearly every design a legislature might try.

3.1 Strict scrutiny and the falsity problem

Reed v. Town of Gilbert (2015) supplies the threshold move: a law is content-based, and therefore subject to strict scrutiny, if it “applies to particular speech because of the topic discussed or the idea or message expressed” (Reed, 2015). Every deepfake-election statute satisfies that definition by design, since each singles out synthetic media about candidates. Strict scrutiny is, in Gerald Gunther’s famous phrase, “strict in theory, fatal in fact” for content-based political-speech restrictions, and the 2026 record bears the aphorism out.

Legislators sometimes assume that falsity supplies an escape hatch: surely the state may ban *false* synthetic depictions of candidates. *United States v. Alvarez* (2012) forecloses that assumption. Striking down the Stolen Valor Act, the Court held that false statements of fact are not a categorically unprotected class of speech; the government may punish falsity only within recognized exceptions such as fraud, defamation, or speech integral to criminal conduct (Alvarez, 2012). A fabricated video of a candidate saying words never spoken is, for First Amendment purposes, much like a fabricated claim about military honors: false, perhaps injurious, but protected unless the statute is drawn within one of the historical exceptions. This is why prohibition statutes drafted around “materially deceptive” media keep collapsing. California’s AB 2839, Hawaii’s Act 191, and Montana’s SB 25 each targeted deceptive synthetic content, and each was held likely or actually unconstitutional, because deception alone, without the elements of defamation or fraud, does not remove speech from First Amendment coverage (Kohls v. Bonta, 2025; Bloomberg Law, 2026; Reuters, 2026b).

Nor does the state’s interest in an informed electorate relax the tailoring demand. Judge Park accepted Hawaii’s interest in electoral integrity as compelling and struck the statute down anyway, finding it not narrowly tailored ([Bloomberg Law, 2026](#)). The recurring judicial intuition, voiced most colorfully by Judge Mendez’s “hammer instead of a scalpel” passage, is that deepfake statutes sweep in parody, satire, and hyperbole alongside genuine deception, and that the state has less restrictive means available ([Vermont Legislature, 2026](#)). Which brings us to the second doctrine.

3.2 Viewpoint discrimination: the Montana innovation

The Montana ruling adds a sharper blade. Most deepfake statutes, including SB 25, punish synthetic media deployed to *injure* a candidate while exempting or ignoring identical synthetic media deployed to *praise* one. Judge Watters treated that asymmetry not as a tailoring defect but as viewpoint discrimination: the law “treats identically deceptive media differently based solely on whether its message is weaponized to damage or deployed to elevate a candidate” ([Reuters, 2026b](#)). Viewpoint discrimination is the rare First Amendment violation that is essentially per se unconstitutional; unlike content-based restrictions, which can in principle survive strict scrutiny, viewpoint-based restrictions almost never do. The doctrinal significance is that the defect is structural, not a matter of drafting care. Any statute that conditions liability on whether the synthetic depiction helps or harms its subject, a natural way to draft a “deceptive deepfake” ban, since praise-posing-as-fact seems less harmful, builds viewpoint discrimination into its elements. The Montana court saw this clearly, and its reasoning threatens every “intent to injure” statute in the thirty-three-state corpus, including the Texas template that other states have copied ([Recording Law, 2026](#)).

3.3 Compelled speech and the disclosure paradox

If prohibition is constitutionally perilous, disclosure would seem the safe harbor. The Supreme Court has repeatedly upheld disclaimer and disclosure requirements in campaign finance, and most states chose that design ([Washington Examiner, 2026](#)). Yet the 2026 record shows disclosure mandates facing First Amendment friction of their own, through the compelled-speech doctrine.

McIntyre v. Ohio Elections Commission (1995) held that compelled identification of the speaker in election literature violates the First Amendment ([McIntyre, 1995](#)). A mandatory “AI-generated” label is compelled speech in the same family: the state forces the speaker to affix a government-prescribed message to political expression. Two courts in the 2026 cycle made exactly this move. Judge Mendez held that AB 2839 impermissibly compelled speech alongside its content-based defects ([Kohls v. Bonta, 2025](#)), and Judge Watters flagged Montana’s mandatory AI disclosure on the same ground, emphasizing that creators must add a government-scripted message rather than merely identify who paid for the communication ([Reason, 2026](#)). Judge Park went further, holding that Hawaii’s mandated disclaimer “would distort the meaning and effect of satirical speech,” converting the disclaimer from a neutral transparency device into a state-imposed alteration of the message’s expressive content ([Bloomberg Law, 2026](#)).

This creates what we call the disclosure paradox. Disclosure is the constitutionally modest alternative to prohibition, but its constitutional footing depends on the disclaimer being genuinely informational rather than expressive, and on the underlying detection question being answerable. A disclaimer regime that cannot reliably determine which speech is synthetic, a problem we document in Section 4, either compels speakers to label their own expression on pain of liability for under-labeling, or compels platforms to label others’ expression, raising the speaker-autonomy concerns *McIntyre* protects. Disclosure is less constitutionally exposed than prohibition, and California’s AB 2355 remains in force, but the 2026 rulings show it is not a First Amendment free zone.

3.4 Prior restraint and the counterspeech alternative

A final doctrinal pressure point concerns remedies. Several statutes, most prominently California’s AB 2655, imposed takedown and reporting duties on platforms, effectively enlisting intermediaries in the prior restraint of political communications. Prior restraints bear “a heavy presumption against [their] constitutional validity” (*NYT v. United States*, 1971). The AB 2655 court avoided the First Amendment question by deciding the case on Section 230 preemption, holding the state mandate preempted as applied to interactive computer service providers (Justia, 2025), but the prior-restraint concern explains why legislatures keep reaching for platform mandates and why courts view them with suspicion: a takedown duty operates before any adjudication of falsity, on political speech, at the speed of the news cycle.

Courts confronting these statutes have repeatedly gestured at the constitutionally preferred alternative: counterspeech. Justice Brandeis’s concurrence in *Whitney v. California* (1927), that “the remedy to be applied is more speech, not enforced silence,” and Justice Holmes’s marketplace-of-ideas dissent in *Abrams v. United States* (1919) supply the canonical formulation (Whitney, 1927; Abrams, 1919). Judge Mendez invoked the same logic in *Kohls*, pointing to fact-checking and responsive speech as less restrictive means (Justia, 2024). The counterspeech answer is often dismissed as naive in an age of synthetic media at scale, and the dismissal has force: counterspeech presumes an audience reachable before the damage is done, and synthetic virality compresses that window. The Russian AI-driven influence operation reported in September 2026, fabricated broadcast footage seeded across social media (DISA, 2026), illustrates the asymmetry: a synthetic claim can circle the globe before any fact-checker has opened the file. But the doctrine matters because strict scrutiny’s narrow-tailoring prong asks precisely whether less restrictive alternatives exist, and as long as courts can name counterspeech, labeling by platforms, and voluntary provenance standards as alternatives, prohibition statutes will struggle to show that suppression is necessary. The lesson of the 2026 cycle is not that counterspeech works; it is that courts will demand it be tried, and that statutes must be drafted to survive that demand. Our framework in Section 6 takes this lesson seriously: rather than dismissing counterspeech as naive, it proposes building the infrastructure, registries, preserved provenance, and researcher access, that would make counterspeech fast enough to matter.

3.5 The doctrinal bottom line

Read together, the four doctrines leave legislators a narrowing corridor. Prohibition of deceptive synthetic political speech must survive strict scrutiny without the benefit of a falsity exception (*Alvarez*), must avoid conditioning liability on whether the depiction helps or harms its subject (*Knudsen*’s viewpoint-discrimination holding), and must prefer less restrictive alternatives such as counterspeech. Disclosure mandates avoid the harshest scrutiny but must survive compelled-speech analysis (*McIntyre*), which turns partly on whether the mandated label distorts expressive content, as Hawaii’s did for satire. Platform takedown duties face preemption under Section 230 and the prior-restraint presumption. The corridor is not empty: carefully drawn disclosure regimes with genuine safe harbors, like California’s surviving AB 2355, may persist. But it is narrow, and it narrows further once the technical premise of disclosure is examined.

A skeptical reader will object that the Montana defect is curable: draft the prohibition without the “intent to injure” element, banning deceptive synthetic depictions of candidates regardless of whether they praise or attack, and the viewpoint-discrimination holding falls away. The objection is correct as far as it goes, but the cure trades one fatal defect for two serious ones. A viewpoint-neutral ban on all deceptive synthetic candidate media must still survive *Alvarez*, which denies that falsity alone removes speech from protection, and it magnifies the overbreadth problem, since a

ban indifferent to valence now also captures flattering synthetic tributes, campaign-approved AI testimonials, and the parody that courts have been most eager to protect. Minnesota’s statute remains in force, and its eventual adjudication on the merits will test whether any prohibition design can thread this needle; as of September 2026, the scorecard reads three injunctions to zero sustained prohibitions, and the doctrinal obstacles are structural rather than drafting errors.

4 The Technical Limits of Detection and Provenance

Disclosure regulation has a hidden technical premise: that AI-generated content can be identified as such, reliably and at scale, so that the duty to disclose can be enforced and the absence of a label can be treated as evidence of human origin. The premise fails. Three families of technical mechanisms, statistical watermarking, cryptographic provenance metadata, and classifier-based detection, each fail in ways that are now documented in the peer-reviewed literature, and each fails most completely in precisely the adversarial conditions that election regulation must contemplate. The point is not that these mechanisms are useless; it is that they are unenforceable at electoral scale against a motivated adversary, which is the only standard that matters for a legal duty.

4.1 Watermark fragility

The most studied provenance mechanism for AI-generated text is statistical watermarking, introduced by [Kirchenbauer et al. \(2023\)](#) in “A Watermark for Large Language Models” (ICML 2023). The scheme biases the model’s token sampling toward a secret “green list,” producing text whose statistical signature a detector can later recognize. In cooperative settings and for long texts, the approach works. The subsequent literature has systematically dismantled its adversarial robustness.

[Sadasivan et al. \(2024\)](#) (“On the Reliability of Watermarks for Large Language Models,” ICLR 2024) demonstrated that paraphrasing, copy-paste, and spoofing attacks degrade watermark detectability, with paraphrase attacks increasing the number of tokens required for confident detection and removing the watermark entirely in short texts. A companion theoretical result showed that as language-model and human text distributions converge, even the best possible detector degrades toward chance ([Sadasivan et al., 2023](#)). [Jovanović et al. \(2024\)](#) (“Watermark Stealing in Large Language Models,” 2024) went further, showing practical spoofing *and* scrubbing attacks against the KGW2-SelfHash scheme previously thought safe: by querying the watermarked API, an adversary can learn the watermark rule and boost scrubbing success from near zero to above 85 percent in hard settings. A 2025 paper on distortion-free watermarks demonstrated spoofing attacks on the KGW, Unigram, and EXP watermark families via public detection APIs ([Breaking Distortion-free Watermarks, 2025](#)).

Two qualifications keep the claim honest. First, Kirchenbauer-style watermarks retain robustness for long texts in non-adversarial settings; the literature does not support the claim that watermarking is categorically impossible. Second, most results concern text; image and audio watermarking face analogous removal attacks but through different mechanisms. The legally relevant claim is narrower and defensible: text watermarks are removable by automated paraphrase and learnable, hence spoofable and scrubbable, by query access, which makes them unenforceable as the technical foundation of a disclosure duty aimed at motivated political actors. A campaign willing to run its synthetic text through a paraphraser, or to query a provider’s detection API to reverse-engineer the watermark, faces no serious technical obstacle, and the cost of doing so is trivial relative to a campaign budget.

4.2 C2PA and the provenance gap

The Coalition for Content Provenance and Authenticity (C2PA), founded in 2021 by Adobe, Arm, the BBC, Intel, and Microsoft, with the Content Authenticity Initiative reporting more than 6,000 members as of January 2026, offers the leading open standard for media provenance: cryptographically signed manifests recording a file’s origin and edit history ([Truescreen, 2026](#)). In principle, a disclosure regime could piggyback on C2PA, requiring AI-generated campaign media to carry signed provenance. In practice, three gaps defeat the design.

First, *stripping is routine*. Most social media platforms recompress and reformat uploaded images, stripping C2PA metadata in the process; as one technical survey puts it, “In most cases, no. The majority of social media platforms recompress and reformat uploaded images, stripping C2PA metadata in the process” ([Truescreen, 2026](#)). A March 2026 survey of generative-media safeguards confirmed that “social media platforms routinely strip C2PA metadata during upload processing,” and a screenshot eliminates provenance entirely ([AI Safeguards Wiki, 2026](#)). The distribution channel that matters most for political persuasion is the channel where provenance is destroyed by default.

Second, *C2PA is not detection*. A manifest records what was declared at signing; a present manifest is not proof the declared information is true, forged manifests with valid signatures are possible, and the absence of a manifest is not evidence that content is synthetic ([Truescreen, 2026](#)). A disclosure duty conditioned on C2PA presence would thus punish compliant actors whose metadata was stripped in transit while missing noncompliant actors entirely.

Third, *adoption is partial and asymmetric*. The same March 2026 survey found only about 38 percent of fifty surveyed generative AI systems implementing adequate machine-readable marking, with 10 percent using invisible watermarks, and reported end-to-end commercial providers near 80 percent marking adoption while API-based services and open-source tools lag far behind ([AI Safeguards Wiki, 2026](#)). A November 2025 audit found that Google and Meta routinely failed to label content their *own* generative tools had created ([MediaNama, 2025](#)), and one technical account reports that only a single phone model signs photos by default while the market-leading handset does not ([Sgoedecke, 2026](#)). (These adoption figures come from industry surveys and technical blogs rather than peer-reviewed studies, and should be read as indicative rather than precise; the directional finding, that marking is uneven and weakest where adversaries operate, is robust.)

The structural point is that C2PA-style provenance is a *cooperative* technology: it works when every handler in the chain preserves the manifest. Electoral adversaries are definitionally non-cooperative, and the dominant distribution platforms strip manifests as a side effect of normal processing. A legal duty that depends on provenance surviving the journey from generator to voter is a duty built on a mechanism that the journey itself destroys.

4.3 Classifier-based detectors

The third mechanism, training classifiers to distinguish AI-generated from human content, fares worst of all. [Liang et al. \(2023\)](#) (“GPT Detectors Are Biased Against Non-Native English Writers,” *Patterns* 2023) tested seven detectors on 91 TOEFL essays and found an average *61.3 percent false-positive rate* on non-native writing, compared with roughly 5 percent on native eighth-grade essays: a detector deployed as an enforcement tool would systematically misclassify the political expression of immigrants and language learners. [Krishna et al. \(2023\)](#) (“Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense,” NeurIPS 2023) showed that automated paraphrasing cut DetectGPT’s accuracy from 70.3 percent to 4.6 percent, an evasion available to anyone with an open paraphrasing model. The RAID benchmark ([Dugan](#)

et al., 2024), the largest open evaluation at 6 million generations across 11 models, 8 domains, and 11 adversarial attacks, concluded that “current detectors are easily fooled by adversarial attacks, variations in sampling strategies, repetition penalties, and unseen generative models.” A 2026 evaluation of commercial detectors found Turnitin at 61 percent and Originality.ai at 69 percent overall accuracy on a balanced benchmark, concluding both are unreliable for high-stakes judgments (Hadra, Cambridge, and Mesbah, 2026).

The detector literature thus converges on a double failure: high false positives against innocent speakers, which makes detector-triggered liability a due-process hazard, and trivial evasion by motivated ones, which makes it useless as an enforcement mechanism. No legislature should predicate legal consequences on a classifier with a 61 percent false-positive rate for non-native writers.

4.4 The open-weight problem

A final structural consideration compounds all three mechanisms. Duties to mark, label, or watermark bind *providers*: commercial entities that can be regulated, inspected, and fined. But openly released model weights, in the class of Llama, Mistral, Qwen, and DeepSeek, can be downloaded and run locally with no provider-side marking, no API logging, and no detection endpoint to query. No peer-reviewed 2026 study was found quantifying this gap, so the claim must be presented as a technical-architecture argument rather than an empirical finding; but the direction is supported by the March 2026 survey finding that open-source tools lag far behind commercial providers in marking adoption (AI Safeguards Wiki, 2026). The consequence is a regulatory asymmetry: marking and disclosure duties fall hardest on the compliant commercial providers whose outputs are already the easiest to identify, while the most capable evaders, running open weights locally, face no technical friction at all. Any provenance regime that does not account for open-weight generation regulates the visible tip of the synthetic-media supply while leaving the submerged bulk untouched.

4.5 What the technical record means for law

The honest summary of the technical literature is not that provenance is impossible. Watermarks work for long cooperative texts; C2PA works within cooperative distribution chains; detectors work on average against naive adversaries. The legally relevant question is different: can these mechanisms support a *legal duty*, enforceable at the scale and speed of a national election, against actors who are trying to evade them? The answer the literature supports is no. Detection-dependent disclosure is technically unenforceable at electoral scale against motivated adversaries. That finding does not end the regulatory inquiry, but it relocates it: if content cannot be reliably identified as synthetic, regulation must attach to something other than content.

Two objections deserve answers. First, the compliance-margin objection: most campaigns are risk-averse, law-abiding actors, and a disclosure duty will change their behavior even if perfect enforcement is impossible. The in-cycle evidence supports this partially: the NRSC’s AI-generated Talarico advertisement did carry a disclaimer, however inconspicuous (Reuters, 2026a). Deterrence at the margin is real, and this paper does not argue disclosure regimes are worthless. The point is distributional: the actors a disclosure duty deters are the established campaigns already subject to reputational and legal discipline, while the actors it cannot reach, foreign influence operations, anonymous PACs, and well-funded practitioners of last-minute synthetic dirty tricks, are precisely where the democratic harm concentrates. A regulatory regime whose enforcement collapses exactly where the threat is greatest is not a regime; it is a compliance tax on the well-behaved. Second, the progress objection: watermarking and detection are improving, and today’s fragility may be

tomorrow’s robustness. Perhaps. But legislation must be grounded in demonstrated capabilities, not anticipated ones, and the trajectory of the literature, from [Kirchenbauer et al. \(2023\)](#) through successive breaking papers, has been an arms race in which each defensive advance is met within months by an offensive one. Law written against the hope that detection will soon work is law written against the evidence.

5 The European Comparison

The European Union’s AI Act (Regulation (EU) 2024/1689) offers the only large-scale alternative to the American state patchwork, and its transparency obligations under Article 50 became applicable on August 2, 2026, just as the American injunction record was taking shape ([European Union, 2024](#); [European Commission, 2026](#)). The comparison is instructive less because Europe solved the problem than because it allocated the duties differently, and the differences illuminate the design error in the American approach.

5.1 Article 50’s duty allocation

Article 50 divides transparency duties between two actors. Article 50(2) obliges *providers* of AI systems that generate synthetic audio, image, video, or text content to ensure that outputs are marked in a machine-readable format and detectable as AI-generated, with the marking required to be effective, interoperable, robust, and reliable ([European Commission, 2026](#); [CSA, 2026a](#)). Article 50(4) obliges *deployers*, the actors who actually use the systems, to disclose when image, audio, or video content constituting a deepfake has been artificially generated or manipulated; a limited exception covers artistic and satirical works, and deployers publishing AI-generated text on matters of public interest must likewise disclose ([CSA, 2026a](#)). Noncompliance carries penalties under Article 99(3) of up to 15 million euros or 3 percent of worldwide annual turnover, whichever is higher ([CSA, 2026a](#)).

Three features of the architecture deserve emphasis. First, the provider-marking duty in Article 50(2) attaches at the point of *generation*, to the entity best positioned to implement it technically, while the deployer-disclosure duty in Article 50(4) attaches at the point of *publication*, to the actor choosing to inject the content into public discourse. The American statutes, by contrast, overwhelmingly load both functions onto the speaker or distributor of the content, demanding that campaigns and platforms detect synthetic provenance that only the provider could reliably have marked. Second, enforcement is *mainly entrusted to national market surveillance authorities*, not to the EU AI Office; the AI Office’s role is limited to AI systems built on general-purpose AI models where the same entity provides both system and model, or systems integrated into VLOPs and VLOSEs designated under the Digital Services Act, with the European Data Protection Supervisor enforcing for EU institutions’ own systems ([European Commission, 2026](#)). American commentary sometimes attributes sweeping enforcement powers to the AI Office; the official allocation is more decentralized. Third, the regime built in a compliance ramp: the Commission adopted guidelines on the Article 50 transparency obligations on July 20, 2026, a voluntary Code of Practice on Transparency of AI-Generated Content offers a lower-scrutiny compliance path, and for AI systems placed on the market before August 2, 2026, the machine-readable-marking obligation for AI-generated content need only be met from December 2, 2026, with content generated before August 2026 not subject to retroactive labeling ([European Commission, 2026](#); [RuleIn, 2026](#)).

The AI Office’s own enforcement machinery over general-purpose AI model providers went live on August 2, 2026, under Articles 88–94 and 101: powers to demand documentation, conduct its own model evaluations, issue corrective orders, restrict or withdraw models, and levy fines up to 3 percent

of global turnover or 15 million euros (CSA, 2026b; Simon Roses, 2026). The underlying GPAI obligations had applied since August 2025; August 2026 switched on the enforcement apparatus, a date the May 2026 Digital Omnibus compromise left intact even as it deferred the high-risk Annex III regime to December 2027 and August 2028 (Regulation AI, 2026).

5.2 What the United States can and cannot borrow

The European model’s virtues are real. By placing the marking duty on providers, it aligns legal obligation with technical capacity: the entity that controls the generation pipeline is the entity that can embed machine-readable provenance at near-zero marginal cost. By placing disclosure on deployers, it targets the actor whose choice creates the democratic harm. And by routing enforcement through market surveillance authorities with technical competence, it avoids asking election officials or generalist courts to adjudicate watermark robustness.

But the European model inherits the same technical ceiling documented in Section 4. Article 50(2)’s requirement that marking be “effective, interoperable, robust, and reliable” is a legal standard awaiting a technical reality that the watermark literature suggests does not yet exist against motivated adversaries (Sadasivan et al., 2024; Jovanović et al., 2024). The open-weight problem bites equally in Europe: a deployer running local open weights faces no provider to compel. And the December 2026 grace period, while sensible as transition policy, means the regime’s first real test will come after the American midterms, so no enforcement record yet exists to evaluate. The United States can borrow Europe’s *allocation logic*, duties at the point of technical capacity rather than at the point of political speech, without borrowing its optimism about the underlying mechanisms. That allocation logic is the foundation of the framework proposed next.

A further European lesson concerns institutional design. The May 2026 Digital Omnibus compromise, which deferred the AI Act’s high-risk Annex III obligations to December 2027 and August 2028 while leaving the August 2026 transparency and GPAI enforcement dates intact, illustrates a legislature sequencing its ambitions: transparency duties first, where the compliance burden is lightest and the technical demands most modest, with the heavier conformity machinery to follow (Regulation AI, 2026). American states have done the reverse, leading with prohibitions, the most constitutionally exposed and technically demanding instrument, while neglecting the infrastructural work, provenance preservation and advertising transparency, that would make any regime administrable. Sequencing matters because each failed prohibition, enjoined within months of enactment, teaches courts to view the next one with suspicion and teaches regulated actors that noncompliance carries little cost. A legislature that began with distributor-side transparency duties would build both the evidentiary record and the compliance culture that harder interventions later require.

6 A Distributor-Duties Framework

The argument so far has been destructive: speaker- and content-focused regulation of synthetic political persuasion is constitutionally imperiled and technically ungrounded. This section proposes a constructive alternative. The core move is to shift the regulatory target from *speakers and content* to *distribution and provenance*: impose duties on the intermediaries that move political media at scale, duties that preserve the information environment’s capacity for self-correction without requiring the state to adjudicate which political claims are synthetic or false.

The framework rests on a simple observation. The state cannot reliably determine whether a given video is AI-generated (Section 4), and it cannot constitutionally punish speakers for the content of their political expression in most cases (Section 3). But the state *can* regulate the

commercial conduct of large distribution platforms, entities already subject to extensive record-keeping, transparency, and consumer-protection obligations, and it can do so through duties that are content-neutral in their operation: they apply to all political media regardless of viewpoint, and they never require a finding that any particular message is false.

6.1 Five duties

1. *Provenance preservation.* Platforms that distribute political advertising above a threshold scale must preserve machine-readable provenance metadata in transit. Where C2PA manifests or equivalent signed metadata accompany uploaded content, recompression and reformatting pipelines must carry the manifests forward rather than stripping them, as current pipelines routinely do (Truescreen, 2026; AI Safeguards Wiki, 2026). This is a commercial-conduct regulation of intermediaries, not a speech restriction: it tells platforms how to handle bits, not what their users may say. Technically it is the cheapest intervention in this paper, since it requires only that platforms stop destroying information they already receive.

2. *Verified-synthetic labeling.* When a platform *knows* content is synthetic, because it carries a valid signed manifest from a provider attesting AI generation, or because the uploader declared it through a political-advertising interface, the platform must label it as such to viewers. Critically, the duty is triggered only by verified provenance, never by a classifier’s suspicion or a government determination. This design sidesteps both the detector-accuracy problem (no liability turns on a 61-percent-false-positive classifier (Liang et al., 2023)) and the compelled-speech problem (the platform labels based on the provider’s or uploader’s own attestation, not on a state-imposed message about contested content). Absence of a label then means only “provenance unverified,” which platforms must communicate honestly rather than implying human origin.

3. *Synthetic political-advertising registries.* Large platforms must maintain public, auditable registries of AI-generated political advertisements: the creative content or a hash thereof, the sponsoring entity, spend, targeting parameters, and impression counts. This extends infrastructure that already exists, the major platforms’ ad libraries, to the synthetic subset, and it operationalizes the counterspeech principle: the remedy for synthetic persuasion is not suppression but sunlight, giving journalists, researchers, and opposing campaigns the data to answer synthetic claims at speed (Whitney, 1927). Because the registry duty attaches to paid distribution rather than to speech content, it parallels long-accepted campaign-finance disclosure without compelling any speaker to alter their message. Implementation can piggyback on existing ad-library APIs: a synthetic-media flag, populated from verified provenance or advertiser self-declaration at upload, plus retention of targeting and reach data for a defined period after the election. The marginal cost to platforms is small relative to the ad-library systems they already operate; the marginal value to election observers is large, since it converts the currently opaque synthetic-ad ecosystem, exemplified by the barely-labeled NRSC advertisement of March 2026 (Reuters, 2026a), into a searchable public record.

4. *Distributor-side accountability.* Liability for breach of the preservation, labeling, and registry duties falls on the distributor, not the speaker. A platform that strips provenance metadata, fails to label content it knows to be synthetic, or omits synthetic ads from its registry faces civil penalties; the speaker whose content was mishandled faces none. This allocation has two virtues. Constitutionally, it avoids strict scrutiny of political speech, since the regulated conduct is the intermediary’s commercial handling of data. Practically, it concentrates enforcement on a small number of sophisticated actors with compliance departments, rather than on thousands of campaigns, PACs, and individual speakers, and it gives platforms a business reason to demand provenance from their upload pipelines.

5. *Safe harbors and scope limits.* The duties apply only above scale thresholds (e.g., platforms

above a monthly-active-user or political-ad-revenue threshold), exempting small forums and individual speakers entirely. Satire and parody distributed through compliant channels receive the same provenance handling as other content, with no additional burden; the framework does not ask anyone to adjudicate humor, the task that most troubled the courts in *Kohls* and *Babylon Bee v. Lopez* (*Kohls v. Bonta*, 2025; *Bloomberg Law*, 2026). And the duties are explicitly without prejudice to Section 230: they regulate platforms’ own conduct as distributors, not their liability for third-party content, avoiding the preemption defect that felled AB 2655 (*Justia*, 2025).

Enforcement design. The duties should be enforceable by state attorneys general and election commissions through civil penalties calibrated to the scale of the violation, not through private rights of action. The caution here is historical: AB 2839’s private right of action invited ideologically motivated litigation against satirists and became part of what the court found objectionable (*Kohls v. Bonta*, 2025). Public enforcement concentrates discretion in accountable officials, allows for graduated remedies (cure periods for first-time preservation failures, escalating fines for registry omissions during the election window), and avoids turning every synthetic meme into a lawsuit. Federal coordination, for instance through the Federal Election Commission’s existing authority over online political advertising disclaimers, could harmonize the registry duty across states, reducing the patchwork compliance burden that the current thirty-three-state regime imposes (*Barclay*, 2026).

6.2 What the framework concedes

Honesty requires stating what distributor duties cannot do. They cannot stop synthetic persuasion that travels through channels outside large platforms: encrypted messaging, email, and offline mailers like the Montana PAC’s AI-generated postcards (*Reuters*, 2026b). They cannot solve the open-weight problem; a locally run model leaves no provider to mark and no platform upload to intercept until the content reaches distribution. And they depend, at the margin, on the same provenance technologies whose limits Section 4 documents: preservation duties are only as good as the manifests that exist to be preserved.

The claim is therefore comparative, not absolute. Against the alternatives actually on the table, prohibition statutes that courts keep enjoining, disclosure mandates whose technical premise fails, and platform takedown duties preempted by federal law, distributor duties are the intervention most likely to survive constitutional review and most closely matched to what the technology can support. They do not require the state to detect synthetic content, to judge political truth, or to silence speakers. They require large intermediaries to stop destroying provenance, to label what they know, and to keep auditable records: modest duties, but ones that compound across the distribution layer where synthetic persuasion actually scales.

The framework’s litigation risk should be stated plainly. A skeptical court could read labeling and registry mandates as treating platforms as publishers of third-party content, inviting the Section 230 preemption analysis that felled AB 2655 (*Justia*, 2025). The distinction we draw is that AB 2655 conscripted platforms into adjudicating and removing users’ speech, the core publisher function, while our duties regulate the platform’s own data handling: preserving bits it receives, labeling based on attestations it holds, and recording transactions it processes. Whether courts accept that line is genuinely uncertain, and any enacted version would likely face an as-applied preemption challenge. Similarly, platform labeling mandates are themselves compelled speech, and their defense would run through the commercial-disclosure doctrine of *Zauderer v. Office of Disciplinary Counsel* (1985) (*Zauderer*, 1985), which permits mandated factual, uncontroversial disclosures: our labeling duty is deliberately designed to satisfy that standard, triggered only by verified provenance and limited to the factual statement that content was AI-generated. These are real vulnerabilities, but they are vulnerabilities of degree, litigated at the margins of commercial regulation, rather than the

head-on collisions with strict scrutiny and viewpoint neutrality that prohibition statutes invite.

7 Related Work

The legal literature on AI-generated election content has grown rapidly. Having proposed a distributor-duties framework above, this section situates it against the contributions closest to it.

[Loewenstein \(2024\)](#), “Make America Fake Again?: Banning Deepfakes of Federal Candidates in Political Advertisements Under the First Amendment” (*Fordham Law Review* 93(1), 2024), reaches the opposite constitutional conclusion from ours: she argues that the proposed federal Protect Elections from Deceptive AI Act would survive First Amendment scrutiny. The disagreement is direct and productive. Loewenstein wrote before the 2026 injunction trilogy, and her analysis accordingly does not confront the viewpoint-discrimination defect identified in *Accountability in State Government v. Knudsen*, which we argue is structural to “intent to injure” statutes, nor does it engage the technical unenforceability literature on which our skepticism about disclosure partly rests. Her argument also leans heavily on the compelling interest in electoral integrity, an interest no court in the 2026 cycle has denied, while underweighting the narrow-tailoring prong where every challenged statute has actually failed. Where she sees a compelling interest sustaining narrowly drawn bans, we see three federal courts in fourteen months finding that interest insufficient to save the statutes actually enacted. The post-midterm record will adjudicate the disagreement further, but as of September 2026 the case law favors our reading.

[Barclay \(2026\)](#), “Regulating Artificial Elections” (SSRN, January 2026), is the closest competitor to this paper’s constructive proposal. Barclay surveys the FEC’s 2024 fraudulent-misrepresentation interpretive rule, state disclaimer laws, and federal proposals, and proposes a hybrid of enforceable AI-content labeling and platform-level, AdChoices-style interventions. We share her turn toward distribution and intermediaries. We differ on two points. First, Barclay’s labeling proposal assumes that AI content can be reliably identified for labeling purposes; the technical record assembled in Section 4 suggests that assumption fails at electoral scale against motivated adversaries, which is why our framework triggers labeling only on verified provenance rather than on detection. Second, Barclay wrote before the Montana injunction and the EU Article 50 guidelines; our synthesis incorporates both, and our provider-versus-deployer allocation analysis draws directly on the European model’s duty split.

A May 2026 survey in the *Journal of High Technology Law* (Suffolk) mapped roughly twenty-eight state laws and the prohibition-versus-disclosure split ([Suffolk JHTL, 2026](#)). It is a useful descriptive baseline, now superseded on the count by Ballotpedia’s thirty-three ([Ballotpedia, 2026](#)), and it does not attempt the doctrinal synthesis of the injunction record or the technical grounding that are this paper’s claimed contributions. Its early-2026 vantage point also predates the developments that most shaped the year’s second half: the permanent California injunctions of August 2025 postdate it only slightly, but the Hawaii ruling of January 2026, the EU Article 50 applicability date, and the Montana preliminary injunction of September 2026 all came later. The survey thus captures the legislative wave at its crest, before the courts began ruling on what the wave had built. The American Enterprise Institute’s April 2026 analysis of *Babylon Bee v. Lopez* treats the Hawaii decision as a cautionary template for other states ([AEI, 2026](#)); we agree, and extend the template into a trilogy with California and Montana, adding the technical-evidence dimension that the AEI analysis, as a doctrinal commentary, does not develop.

The paper’s claimed novelty is thus threefold: (i) the first synthesis of the 2026 injunction trilogy, California’s permanent injunctions of 2025, Hawaii’s January 2026 ruling, and Montana’s September 2026 preliminary injunction, as a single doctrinal pattern rather than isolated setbacks;

(ii) a regulatory argument grounded in the current technical evidence on provenance fragility rather than in speculation about future capabilities; and (iii) the diagnosis of a provider-marking versus deployer-disclosure allocation error in American legislative discourse, informed by the EU Article 50 model that entered application in August 2026.

8 Conclusion

The campaign preceding the November 2026 midterms has been the first in which generative AI functioned as routine campaign infrastructure, and the legal system’s response has been a fast-moving natural experiment. Thirty-three states enacted political-deepfake laws; federal courts enjoined the most aggressive of them on First Amendment, compelled-speech, viewpoint-discrimination, and Section 230 preemption grounds; and the technical mechanisms on which the surviving disclosure regimes depend have been shown, in the peer-reviewed literature, to fail against the adversaries they must restrain. The experiment’s lesson is not that regulation is futile but that it has been aimed at the wrong target. Speakers cannot constitutionally be made to label or suppress synthetic political speech in most cases, and synthetic speech cannot technically be detected at the scale and speed elections require. What can be regulated, constitutionally and technically, is distribution: the preservation of provenance by large intermediaries, the labeling of content verifiably known to be synthetic, and auditable public records of synthetic political advertising.

For legislators drafting the post-midterm agenda, the recommendations follow directly. Stop writing statutes that condition liability on whether synthetic media helps or harms a candidate; the Montana court’s viewpoint-discrimination analysis will find them. Stop predicating legal duties on detector outputs or on the assumption that watermarks survive adversarial handling; the literature will not support the enforcement actions that follow. Start allocating duties to the point of technical capacity, providers marking at generation, platforms preserving in transit, as the EU’s Article 50 model attempts. And start building the evidentiary infrastructure, registries, preserved provenance, researcher access, that makes counterspeech fast enough to matter. The First Amendment, correctly understood, does not forbid election-integrity regulation; it forbids only regulation that suppresses speech the state cannot reliably identify and constitutionally cannot punish. A distributor-duties framework works within that boundary. Whether legislatures will accept its modesty, trading the satisfaction of banning deepfakes for the unglamorous work of preserving metadata, is a political question this paper cannot answer. The technical and constitutional analysis suggests it is the only question worth asking.

References

Abrams v. United States, 250 U.S. 616 (1919) (Holmes, J., dissenting).

American Enterprise Institute (2026). Regulating Political Deepfakes: First Amendment Lessons from Hawaii’s Unconstitutional Effort. <https://www.aei.org/technology-and-innovation/regulating-political-deepfakes-first-amendment-lessons-from-hawaiis-unconstitutional-effort/>

AI Safeguards Wiki (2026). C07-07: Generative Media Safeguards: survey of generative AI systems’ marking adoption (March 2026). <https://github.com/preechapon250/aisvs/blob/HEAD/wiki/C07-07-Generative-Media-Safeguards.md>

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

- Ballotpedia (2026). State of Deepfake Legislation 2026 Annual Report. Announced Aug. 5, 2026 via Newsfile Corp.; Pennsylvania tracker: “As of September 3, 2026, 33 states had enacted laws regulating deepfakes used in political communications.” https://ballotpedia.org/AI_deepfake_policy_in_Pennsylvania
- Barclay, C. A. (2026). Regulating Artificial Elections. SSRN, posted Jan. 14, 2026. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6069453
- Bloomberg Law (2025). X’s Preemption Claim on Minnesota Political Deepfakes Law Fails (Dec. 2025). <https://news.bloomberglaw.com/litigation/xs-preemption-claim-on-minnesota-political-deepfakes-law-fails>
- Bloomberg Law (2026). Hawaii’s Deepfake Election Law Violates Free Speech, Court Finds (Feb. 2026). Reporting on *The Babylon Bee LLC v. Lopez*, D. Haw., permanent injunction Jan. 30, 2026 (Park, J.). <https://news.bloomberglaw.com/litigation/hawaiis-deepfake-election-law-violates-free-speech-court-finds>
- Citizen Portal (2026). States Split on Deepfakes, NCSL Warns Many Laws Face First Amendment Tests (2026). <https://citizenportal.ai/articles/7461166/nevada/2026-legislature-nv/nevada/2026-legislature-nv/nevada/2026-legislature-nv/Nevada/States-split-on-deepfakes-NCSL-warns-many-laws-face-First-Amendment-tests>
- Cloud Security Alliance (2026). Research Note: EU AI Act Article 50 Transparency (July 29, 2026). <https://labs.cloudsecurityalliance.org/research/csa-research-note-eu-ai-act-article-50-transparency-20260729/>
- Cloud Security Alliance (2026). Research Note: EU AI Act GPAI Enforcement (May 9, 2026). <https://labs.cloudsecurityalliance.org/research/csa-research-note-eu-ai-act-gpai-enforcement-20260509-csa-st/>
- DISA (2026). U.S. Intelligence Warns Russia Is Mounting an AI-Driven Disinformation Campaign Ahead of Elections (c. Sept. 18, 2026). <https://disa.org/u-s-intelligence-warns-russia-is-mounting-an-ai-driven-disinformation-campaign-ahead-of-elections/>
- Dugan, et al. (2024). RAID: the largest open benchmark of AI-text detectors (6M generations, 11 models, 8 domains, 11 adversarial attacks). Cited in https://github.com/kitfoxs/humanize-mcp/blob/HEAD/research/00_executive_summary.md
- ElectionDesk (2026). Federal Court Strikes Down Hawaii Deepfake Law on First Amendment Grounds (Feb. 2, 2026). <https://electiondesk.org/2026/02/02/federal-court-strikes-down-hawaii-deepfake-law-on-first-amendment-grounds/>
- European Commission (2026). Transparency Obligations under Article 50 of the AI Act. Article 50 applicable as from 2 August 2026; enforcement mainly by national market surveillance authorities; grace period to 2 December 2026 for pre-existing systems’ marking duties. <https://digital-strategy.ec.europa.eu/en/node/17084/printable/pdf>
- European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Federal Communications Commission (2024). Declaratory Ruling banning AI-generated voices in robocalls under the Telephone Consumer Protection Act (Feb. 8, 2024).

- Agence France-Presse (2026). AI Political Ads Warp Reality Ahead of US Midterms (Sept. 22, 2026). <https://www.lagacetademexico.mx/en/Economy/745996-ai-political-ads-warp-reality-ahead-of-us-midterms.html>
- Hadra, Cambridge, and Mesbah (2026). Evaluation of commercial AI-text detectors (Turnitin 61%, Originality.ai 69% overall accuracy on a 192-text balanced benchmark). Cited in <https://deceptioner.site/blog/ways-to-spot-ai-writing-in-classrooms>
- Jovanović, et al. (2024). Watermark Stealing in Large Language Models. arXiv:2402.19361.
- Justia (2024). Docket: *Kohls v. Bonta*, No. 2:24-cv-02527-JAM-CKD (E.D. Cal.). <https://docs.justia.com/cases/federal/district-courts/california/caedce/2%3A2024cv02527/453046/101>
- Justia (2025). Order permanently enjoining Cal. AB 2655 on 47 U.S.C. § 230 preemption grounds (Aug. 20, 2025), consolidated *Kohls* docket. <https://cases.justia.com/federal/district-courts/california/caedce/2:2024cv02527/453046/98/0.pdf?ts>
- Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A Watermark for Large Language Models. In *Proc. 40th Int’l Conf. on Machine Learning (ICML)*. arXiv:2301.10226.
- Kohls v. Bonta*, 797 F. Supp. 3d 1177 (E.D. Cal. 2025) (Mendez, J.). Permanent injunction of Cal. AB 2839 on First Amendment grounds (Aug. 29, 2025); preliminary injunction Oct. 2, 2024.
- Krishna, et al. (2023). Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- LegalNewsFeed (2026). Minnesota AG Seeks Dismissal of X Corp’s Challenge to State Election Deepfake Law (Sept. 21, 2026). <https://legalnewsfeed.com/2026/09/21/minnesota-ag-seeks-dismissal-of-x-corp-s-challenge-to-state-election-deepfake-law/>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT Detectors Are Biased Against Non-Native English Writers. *Patterns* 4(7), 100779. arXiv:2304.02819.
- Loewenstein, S. (2024). Make America Fake Again?: Banning Deepfakes of Federal Candidates in Political Advertisements Under the First Amendment. *Fordham Law Review* 93(1), 273. <https://ir.lawnet.fordham.edu/flr/vol93/iss1/7/>
- McIntyre v. Ohio Elections Comm’n*, 514 U.S. 334 (1995).
- MediaNama (2025). Audit: Social Media, Google, Meta AI Labelling (Nov. 2025). <https://www.medianama.com/2025/11/223-audit-social-media-google-meta-ai-labelling/>
- Mint (2026). AI Deepfakes Are Getting Weirder and Harder to Spot in the Midterms (June 15, 2026; via Wall Street Journal). <https://www.livemint.com/global/ai-deepfakes-are-getting-weirder-and-harder-to-spot-in-the-midterms-11781514477398.html>
- New York Times Co. v. United States*, 403 U.S. 713 (1971).
- Reason (2026). Montana’s Anti-Deepfakes Law Just Hit a First Amendment Roadblock (Sept. 18, 2026). <https://reason.com/2026/09/18/montanas-anti-deepfakes-law-just-hit-a-first-amendment-roadblock/>

Recording Law (2026). US Deepfake Laws Tracker. <https://www.recordinglaw.com/us-laws/deepfake-laws/>

Reed v. Town of Gilbert, 576 U.S. 155 (2015).

Regulation AI (2026). EU AI Act Enforcement August 2026. <https://www.regulation-ai.eu/en/blog/eu-ai-act-enforcement-august-2026/>

Ax, J. and Coster, H. (2026). AI Deepfakes Blur Reality in 2026 US Midterm Campaigns. Reuters, Mar. 28, 2026. <https://www.reuters.com/business/media-telecom/ai-deepfakes-blur-reality-2026-us-midterm-campaigns-2026-03-28/>

Reuters (2026). US Judge Blocks Montana Enforcing Deepfake Election Ads Law Against Conservative. Sept. 17, 2026. Reporting on *Accountability in State Government v. Knudsen*, D. Mont. (Watters, J.), preliminary injunction Sept. 17, 2026. <https://www.reuters.com/legal/government/us-judge-blocks-montana-enforcing-deepfake-election-ads-law-against-conservative-2026-09-17/>

Simon Roses (2026). The Day the AI Act Grew Teeth: GPAI Enforcement Goes Live (blog, Aug. 2026). <https://simonroses.com/2026/08/the-day-the-ai-act-grew-teeth-gpai-enforcement-goes-live/>

RuleIn (2026). Summary: Commission Guidelines on AI Act Transparency Obligations (adopted July 20, 2026). https://www.rulein.eu/web/content/blog.post/9/rulein_pdf_attachments?download=true&filename=AIActTransparencyGuidelines_EN.docx.pdf

Sadasivan, et al. (2023). Can AI-Generated Text be Reliably Detected? arXiv:2303.11156.

Sadasivan, et al. (2024). On the Reliability of Watermarks for Large Language Models. In *Proc. Int'l Conf. on Learning Representations (ICLR)*. arXiv:2306.04634.

Sgoeddecke, S. (2026). C2PA Only Works If Everything Is Signed (technical blog). <https://github.com/sgoeddecke/gatsby-blog/blob/HEAD/content/blog/c2pa-only-works-if-everything-is-signed/index.md>

Journal of High Technology Law (2026). Deepfakes at the Ballot Box: The Challenge of Regulating AI in Political Campaigns (May 9, 2026). Suffolk University. <https://sites.suffolk.edu/jhtl/2026/05/09/deepfakes-at-the-ballot-box-the-challenge-of-regulating-ai-in-political-campaigns/>

Truescreen (2026). C2PA Standard: History and Limitations. <https://truescreen.io/articles/c2pa-standard-history-limitations/>

Trump v. CASA, 600 U.S. ____ (2025) (limiting universal injunctions).

Vermont Legislature (2026). Synopsis of *Kohls v. Bonta*, 797 F. Supp. 3d 1177 (2025), prepared for Senate Government Operations (Jan. 16, 2026). <https://legislature.vermont.gov/Documents/2026/Workgroups/SenateGovernmentOperations/Bills/S.23/Drafts,Amendments,andLegalDocuments/S.23~RikSehgal-KohlsvBonta~1-16-2026.pdf>

Washington Examiner (2026). AI Deepfake Election Advertising Laws: What to Know (c. Aug. 2026). <https://www.washingtonexaminer.com/news/campaigns/4680277/ai-deepfake-election-advertising-laws-what-to-know/>

Whitney v. California, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring).

Zauderer v. Office of Disciplinary Counsel of the Supreme Court of Ohio, 471 U.S. 626 (1985).

Breaking Distortion-free Watermarks in Large Language Models (2025). arXiv:2502.18608.