Revisiting Defamation Laws in the Age of AI-Generated Libel

ABBY POWNALL[†]

Rapid advancements in artificial intelligence (AI) pose an emerging crisis for the spread of political misinformation. Generative AI, capable of producing hyper-realistic deepfakes, synthetic audio, and fabricated video, enables private individuals to construct and disseminate defamatory content at an unprecedented scale and speed. When libel against public officials is generated using these tools, the actual malice standard is satisfied as the user cannot produce such an output without knowingly and intentionally prompting the AI system to fabricate representations that appear factual. Though defendants may invoke the satire defense, the increasing realism of deepfakes has blurred the line between satire and factual assertions. The psychological impact of images and videos exceeds that of text, rendering the reasonable person standard traditionally applied in cases of satire unstable, and ultimately calling into question courts' ability to rely on that standard in cases involving synthetic media. This Note argues that the Supreme Court's landmark decision in New York Times Co. v. Sullivan rests on empirical and psychological assumptions about how speech circulates, how reputational harm can be corrected, and how audiences distinguish fact from satire—assumptions that no longer hold in an information ecosystem dominated by generative AI and synthetic media. Sullivan remains a pillar of American free speech jurisprudence, and while its normative commitment to uninhibited debate remains sound, its doctrinal architecture must be recalibrated to account for how AI-generated content undermines the feasibility of counter-speech, the effectiveness of retractions, and the workability of the satire defense and the reasonable person doctrine.

[†] J.D. Candidate 2026, University of California, College of the Law, San Francisco. I am deeply grateful to Executive Notes Editor Madeleine Heldman, the Notes team, and staff editors of *UC Law Journal* for their rigorous review process and contributions to bettering this Note. My sincere thanks to Professor Paul Belonick for encouraging me to pursue this topic and for his thoughtful feedback throughout my writing process. Finally, I would like to thank my parents, brothers, and friends for their unwavering belief in me and constant support and encouragement throughout law school.

TABLE OF CONTENTS

INTRODUCTION 1441
I. THE LEGAL LANDSCAPE OF DEFAMATION LAWS 1442
A. CURRENT DEFAMATION LAWS 1444
B. HEIGHTENED STANDARD FOR PUBLIC OFFICIALS: ACTUAL MALICE 1444
1. The Policy Behind Actual Malice 1446
a. Heightened Tolerance as Public Officials 1446
b. Uninhibited Debate 1446
c. Self-censorship 1447
C. ASSUMPTIONS UNDERLYING CURRENT DEFAMATION LAWS 1448
1. The Ability to Combat 1448
2. The Ability to Retract 1448
3. The Ability to Get the Joke 1449
a. The Satire Defense 1449
II. THE RIDE OF AI LIBEL 1450
A. GENERATIVE AI IN TODAY'S POLITICAL ARENA 1451
B. *WALTERS V. OPENAI, L.L.C.*—THE FIRST OF MANY? 1452
C. ASSUMPTIONS UNDERLYING TODAY'S REALITY 1454
1. The Ability to Combat and Retract, Undermined by Virality 1455
2. Satire and the Collapse of the Reasonable Person Standard 1456
a. The Satire Defense 1458
III. REFORMS 1459
A. CURRENT ATTEMPTS AT CONTAINING AI-LIBEL 1460
B. WHY EXISTING REFORMS FAIL IN AN INTERNET-DRIVEN ECOSYSTEM 1461
1. Constitutional Limits on Regulating Private AI Use 1461
2. The Attribution Problem: Difficulty in Identifying the Speaker 1462
3. Liability Fractures Across Platforms, Users and Algorithms 1462
C. PRESERVING THE MARKETPLACE OF IDEAS IN AN AI-DISTORTED ENVIRONMENT 1463
CONCLUSION 1464

INTRODUCTION

The integration of artificial intelligence (AI) into everyday communication has dramatically altered how ideas and information are produced and disseminated. One consequence of this is the unprecedented ease with which disinformation can spread and the unfounded complexities for First Amendment jurisprudence. Generative AI uses a large language model (LLM) to create new, original content from existing patterns and data stored within the LLM.[1] These models rely on massive datasets and refine their outputs through iterative learning, increasing their breadth of knowledge over time.[2] LLM outputs are not limited to text, but also include images, music, and video.[3] Among the most consequential of these outputs are "deepfakes"—"fabricated hyper-realistic" videos, images, and audio.[4] Such creations have been frequently "altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said."[5] Such "deepfake technology" is capable of manipulating facial expressions, synthesizing faces, and fabricating speech by leveraging existing data about an individual's voice and appearance.[6] The central concern with deepfake technology is the heightened risk that the viewers may be unable to decipher fiction from reality.

Despite common law tort doctrines, heightened constitutional protections for political speech, and the demanding actual malice standard for public figures, the law struggles to account for these rapid technological developments. As a result, individuals who generate AI-based libel are increasingly insulated from liability, while the spread of such content becomes nearly impossible to contain.[7]

These developments expose a deeper doctrinal problem. Modern defamation law, particularly as applied to public officials, rests on a set of unstated factual and psychological assumptions about speech that no longer adequately reflect how information is created, perceived, or corrected in an AI-mediated information environment. At the center of this tension lies *New York Times Co. v. Sullivan*, the Supreme Court's landmark decision establishing the

---

1. *Generative Artificial Intelligence*, CORN. UNIV.: CTR. FOR TEACHING INNOVATION, https://teaching.cornell.edu/generative-artificial-intelligence (last visited Apr. 11, 2026).

2. *Id.*

3. *Id.*

4. *Dangers of Deepfake: What to Watch For*, STAN. UNIV. IT: NEWSROOM (Feb. 22, 2024), https://uit.stanford.edu/news/dangers-deepfake-what-watch.

5. *Deepfake*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/deepfake (last visited Apr. 11, 2026).

6. TIMOTHY M. PERSONS, U.S. GOV'T ACCOUNTABILITY OFF., GAO-20-379SP, DEEPFAKES (2020) ("Deepfakes rely on artificial neural networks, which are computer systems modeled loosely on the human brain that recognize patterns in data. Developing a deepfake photo or video typically involves feeding hundreds or thousands of images into the artificial neural network, 'training' it to identify and reconstruct patterns—usually faces.").

7. Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza & Hafiz Malik, *Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward*, 53 APPLIED INTEL. 3974, 3974–75 (2023).

actual malice standard for defamation claims brought by public officials.[8] Under that standard, a public official may recover for defamation only by proving that the challenged statement was made with knowledge of its falsity or with reckless disregard for whether it was true or false.[9] *Sullivan* did not merely articulate a heightened fault requirement. It also relied on a set of background assumptions about the structure of public discourse. According to the Court, public officials possess meaningful access to counter-speech, false statements can be feasibly retracted or corrected through established media channels, and audiences can distinguish fact from satire, parody, or exaggeration.[10]

This Note argues that generative AI, especially deepfake technology, systemically undermines each of the Court's assumptions. Unlike the media environment in which *Sullivan* was decided, today's information ecosystem is characterized by algorithmic amplification, viral dissemination, and synthetic media that can appear indistinguishable from reality.[11] As a result, falsehoods spread faster than rebuttals, retractions are functionally ineffective, and the traditional reasonable person standard—long relied upon to distinguish satire from actionable defamation—loses coherence when confronted with hyper-realistic falsity.

I. THE LEGAL LANDSCAPE OF DEFAMATION LAWS

Until the landmark decision of *Sullivan*, defamatory speech was largely unprotected under the First Amendment,[12] as regulation of defamatory speech was left almost entirely to the states.[13] Prior to *Sullivan*, a plaintiff who sued for libel did not need to prove the statement's falsity or demonstrate actual injury.[14] The plaintiff only needed to show that harm was a possible consequence of the statement.[15] Though truth was still a complete defense to libel, it was not easy to establish because the strict standard required the statement to be proved true "in all its particulars."[16] The ease with which plaintiffs could succeed in

8. 376 U.S. 254, 279–80 (1964).

9. *Id.*

10. *See, e.g.*, *id.* at 286–87 (discussing broadly the ability of newspapers to retract); *id.* at 304–5 (Goldberg, J., concurring) (noting public officials' greater access to media of communication than that of most private citizens and the ability of counterargument to expose misstatements); *id.* at 270–271, 279 n. 19 (recognizing that public discourse may include exaggeration, caustic language, and rhetorical flourish that audiences are expected to contextualize rather than treat as literal factual assertions).

11. *See, e.g.*, Donghee Shin & Emily Y. Shin, *Cascading Falsehoods: Mapping the Diffusion of Misinformation in Algorithmic Environments*, 41 AI & SOC'Y 1435, 1437 (2026); Soroush Vosoughi, Deb Roy & Sinan Aral, *The Spread of True and False News Online*, 359 SCI. 1146, 1146 (2018).

12. Samantha Barbas, *The Enduring Significance of* New York Times Co. v. Sullivan, KNIGHT FIRST AMEND. INST. AT COLUM. UNIV.: BLOG (Mar. 18, 2024), https://knightcolumbia.org/blog/the-enduring-significance-of-new-york-times-v-sullivan; *Sullivan*, 376 U.S. at 256.

13. Barbas, *supra* note 12; *see* Chaplinksy v. New Hampshire, 315 U.S. 568, 571–73 (1942); Beauharnais v. Illinois, 343 U.S. 250, 265–67 (1952).

14. Barbas, *supra* note 12. Prior to *Sullivan*, falsity was presumed. *Id.*

15. *Id.*

16. *Id.*

defamation lawsuits consequently deterred news outlets from freely publishing criticisms of public officials for fear of legal reprisal.[17]

Since *Sullivan*,[18] certain political leaders, like President Donald Trump, have attacked the actual malice standard, arguing that libel laws should "open up."[19] Supreme Court Justices have also suggested that the landmark decision should be reconsidered.[20] In particular, Justice Clarence Thomas has been a longtime critic of the actual malice standard, arguing that it has no basis in the Constitution.[21] In *Blankenship v. NBC Universal, LLC*, Thomas concurred in the denial of the petition for a writ of certiorari as warranted as a matter of state law, though he nonetheless urged the Court to reconsider *Sullivan* when an appropriate case arose.[22] In his concurrence, Thomas denounced the *Sullivan* Court's decision to "allow[ ] media organizations and interest groups 'to cast false aspersions on public figures with near impunity'" and urged the Court to reconsider *Sullivan*.[23] Justice Neil M. Gorsuch has also expressed skepticism regarding the doctrine's continued viability.[24] A wholesale rollback of *Sullivan* would substantially alter the balance between reputational protections and political speech, raising serious First Amendment concerns.[25] This would threaten the First Amendment's foundational premise: For truth to prevail, even pernicious opinions must enter the marketplace of ideas. The press and its ability to criticize freely without fear of consequences play a critical role in upholding these values.

Calls for *Sullivan* to be reconsidered, however, are not entirely unfounded. Looking at the way that sources of news have evolved concurrently with technological evolutions, the news media landscape in 1964 is fundamentally distinct from the modern news landscape. Rather than having a few dominant

17. Stephen M. Renas, Charles J. Hartmann & James L. Walker, *An Empirical Analysis of the Chilling Effect: Are Newspapers Affected by Liability Standards in Defamation Actions?*, *in* THE COST OF LIBEL: ECONOMIC AND POLICY IMPLICATIONS 41, 43 (Everette E. Dennis & Eli M. Noam eds., 1989).

18. See *infra* Subpart.I.B for more background on the *Sullivan* decision.

19. Michael M. Grynbaum, *Trump Renews Pledge to 'Take a Strong Look' at Libel Laws*, N.Y. TIMES (Jan. 10, 2018), https://www.nytimes.com/2018/01/10/business/media/trump-libel-laws.html.

20. Adam Liptak, *Two Justices Say Supreme Court Should Reconsider Landmark Libel Decision*, N.Y. TIMES (July 2, 2021), https://www.nytimes.com/2021/07/02/us/supreme-court-libel.html.

21. Blankenship v. NBCUniversal, LLC, 144 S. Ct. 5, 5 (2023) (Thomas, J., concurring) ("The Court did not base this 'actual malice' rule in the original meaning of the First Amendment. . . . '*New York Times* and the Court's decisions extending [the actual malice standard] were policy-driven decisions masquerading as constitutional law.' The decisions have 'no relation to the text, history, or structure of the Constitution.'") (emphasis omitted) (citations omitted) (quoting McKee v. Cosby, 139 S. Ct. 675, 676 (2019) (Thomas, J., concurring); Tah v. Glob. Witness Publ'g, 991 F.3d 231, 251 (D.C. Cir. 2021) (Silberman, J., dissenting)).

22. *Id.*

23. *Id.* (citing *Tah*, 991 F.3d at 254).

24. Berisha v. Lawson, 141 S. Ct. 2424, 2428 (2021) (Gorsuch, J., dissenting) ("What started in 1964 with a decision to tolerate the occasional falsehood to ensure robust reporting by a comparative handful of print and broadcast outlets has evolved into an ironclad subsidy for the publication of falsehoods by means and on a scale previously unimaginable.").

25. *See* Samantha Barbas, *Overturning* New York Times vs. Sullivan *Would Endanger Freedom of Speech*, U.C. PRESS BLOG (Dec. 20, 2022), https://www.ucpress.edu/blog-posts/61144-overturning-new-york-times-vs-sullivan-would-endanger-freedom-of-speech.

reputable news sources, there are now dozens of traditional news outlets alongside less traditional channels on social media.[26] Libelous statements have also transformed in medium with an increasing presence of AI-generated falsehoods. This Note examines the *Sullivan* standard, the policy considerations underlying the Court's reasoning, and the modern policy concerns that warrant renewed scrutiny today.

A. CURRENT DEFAMATION LAWS

The First Amendment establishes that "Congress shall make no law . . . abridging the freedom of speech . . . ."[27] Libel is defined as a form of defamation expressed in written or other fixed forms of communication that "is injurious to a person's reputation; exposes a person to public hatred, contempt or ridicule; or injures a person in their business or profession."[28] Though some states vary slightly, to prevail on a defamation claim, the plaintiff bears the sole burden of establishing four elements: "(1) a false statement purporting to be fact; (2) publication or communication of that statement to a third person; (3) fault amounting to at least negligence; and (4) damages, or some harm caused to the reputation of the person or entity who is the subject of the statement."[29] Defamation encompasses both libel (written) and slander (spoken).[30] Given the integration of the internet into everyday interactions, an increasing amount of defamation stems from online publications, or "cyber-libel."[31] AI-generated defamatory statements—particularly those conveyed through images, audio, and video—fall squarely within the category of libel.

B. HEIGHTENED STANDARD FOR PUBLIC OFFICIALS: ACTUAL MALICE

When it comes to libel against political figures, the standards are heightened compared to those governing libel against private individuals.[32]

---

26. *See Social Media and News Fact Sheet*, PEW RSCH. CTR. (Sep. 25, 2025), https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet (noting the use of social medias such as Instagram, TikTok, or X as places where Americans now regularly get news).

27. U.S. CONST. amend. I. This protection is further incorporated against the states through the Fourteenth Amendment. Gitlow v. New York, 268 U.S. 652, 666 (1925); N.Y. Times Co. v. Sullivan, 376 U.S. 254, 276–77 (1964).

28. *Libel*, CORN. L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/libel (last visited Apr. 12, 2026).

29. *Defamation*, CORN. L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/defamation (last visited Apr. 12, 2026).

30. RESTATEMENT (SECOND) OF TORTS §§ 558, 568 (A.L.I. 1965).

31. Barry J. Waldman, Comment, *A Unified Approach to Cyber-Libel: Defamation on the Internet, a Suggested Approach*, 6 RICH. J.L. & TECH., Fall 1999, at 1, 1.

32. *See generally* Gertz v. Robert Welch, 418 U.S. 323 (1974). In this case, the Court held that states may impose laws with a lower burden of proof for private individuals than public officials in defamation cases. Rather than actual malice, states need only require private individuals to prove fault and actual damages to prevail in a defamation claim.

Sullivan established the heightened standard for "actual malice" in cases involving "public officials" or "matters of public concern."[33]

In 1960, The New York Times published an advertisement, paid for by civil rights activists, that solicited donations to defend Martin Luther King, Jr. on perjury charges.[34] The advertisement also criticized the Montgomery, Alabama police department and its response to the widespread civil rights demonstrations.[35] However, the advertisement contained several minor factual inaccuracies when describing some of the events that occurred in Montgomery.[36] For example, the advertisement misstated the names of songs that protesters sang, suggested that students were expelled for leading a demonstration at the state capitol when they had in fact been disciplined for a different protest, and inaccurately reported that police had padlocked the local college's dining hall.[37] L.B. Sullivan, Montgomery's Commissioner of Public Affairs, claimed that his reputation was damaged by the criticisms of his subordinates and the advertisement's inaccuracies, even though Sullivan was not mentioned by name.[38] After sending a request for The New York Times to publicly retract the information, the newspaper refused.[39] In response, Sullivan filed a libel action.[40] The issue before the Court was whether a publication forfeits its constitutionally protected ability to openly criticize a public official by including factual inaccuracies. The Court ruled in favor of The New York Times and established the actual malice standard.

In *Curtis Publishing Co. v. Butts*, the Court expanded the category of plaintiffs subject to the actual malice standard by applying it not only to public officials, but also to public figures bringing defamation claims.[41] The Court there reasoned that public figures, like public officials, often possess access to effective channels of communication and voluntarily assume positions of prominence that invite public scrutiny, rendering speech about them similarly central to First Amendment concerns.[42] Thus, the rationale turned less on government office and more on the nature of public participation and public scrutiny. *Sullivan*'s actual malice standard adds a heightened requirement to a standard defamation claim. Rather a negligence-based standard, the plaintiff must demonstrate that the defamatory statement was made with either (1) knowledge that the statement was false or (2) reckless disregard for whether the statement was true or false.[43] Additionally, the standard of proof is heightened

33. N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–80 (1964).
34. *Id.* at 256–57.
35. *Id.* at 257–58.
36. *Id.* at 258.
37. *Id.* at 258–59.
38. *Id.* at 258.
39. *Id.* at 261.
40. *Id.*
41. Curtis Pub. Co. v. Butts, 388 U.S. 130, 154–55 (1967).
42. *Id.*
43. *Sullivan*, 376 U.S. at 280–81.

from a "preponderance of the evidence" to a "clear and convincing" standard.[44] The Court in *Sullivan* explained that such claims require "the convincing clarity which the constitutional standard demands . . . ."[45]

1. The Policy Behind Actual Malice

In *Sullivan*, the Court reasoned that this higher legal standard is necessary to strike an appropriate balance between the constitutional protection of free speech and the press, and the reputational interests of public officials.[46] The overarching concern was that less demanding defamation laws pose a threat to First Amendment freedoms by chilling political speech.[47] The decision was grounded in the institutional role of public officials and the need to preserve uninhibited public debate on matters of public concern.

a. Heightened Tolerance as Public Officials

The *Sullivan* decision makes successful libel litigation exceptionally difficult for public official plaintiffs. However, this result has been justified by the heightened level of criticism that public officials agree to when accepting such a role.[48] Due to the power imbalance between public officials and their constituents, the Court feared that public officials might deploy libel laws as a tool to suppress dissenting opinions.[49] Without heightened constitutional protections, defamation liability could operate as a form of indirect government censorship by deterring criticisms of official conduct through the threat of civil damages.[50] Given their assumed public role and corresponding influence over public policy and public opinion, the Court concluded that public officials are subject to a heightened level of scrutiny and are expected to tolerate a broader range of criticism without recourse to defamation claims.[51]

b. Uninhibited Debate

The Court grounded its reasoning in the nation's longstanding commitment to safeguarding "uninhibited, robust, and wide-open" debate on issues of public concern, even when such debate includes "sometimes unpleasantly sharp attacks" on public officials.[52] This reasoning reflects the "marketplace of ideas" theory, which posits that truth can only emerge from the free competition of

44. *Id.* at 285–86. *Sullivan* first introduced the "convincing clarity" language, which was later clarified as a clear-and-convincing evidentiary burden. Bose Corp. v. Consumers Union of U.S., Inc., 466 U.S. 485, 511–12 (1984).

45. *Sullivan*, 376 U.S. at 285–86.

46. *Id.* at 281.

47. *Id.* at 270, 282.

48. *Id.* at 270.

49. *Id.* at 277 ("What a State may not constitutionally bring about by means of a criminal statute is likewise beyond the reach of its civil law of libel.").

50. *Id.*

51. *Id.* at 273.

52. *Id.* at 270.

ideas in the public square.[53] As Judge Learned Hand famously observed, the First Amendment "presupposes that right conclusions are more likely to be gathered out of a multitude of tongues, than through any kind of authoritative selection."[54] *Sullivan* further emphasized the press's essential role in facilitating uninhibited debate and promoting accountability among those in power, explaining that the press exists not to shield the reputations of the powerful but to inform the public.[55] At the same time, the Court acknowledged that individuals, including public officials, retain legitimate interests in protecting their reputations.[56] Nevertheless, *Sullivan*'s actual malice standard prioritizes First Amendment freedoms over reputational interests, reflecting the Court's judgment that occasional false or exaggerated statements are an acceptable cost of preserving robust political discourse.[57]

c. Self-censorship

The First Amendment was "fashioned to assure unfettered interchange of ideas for the bringing about of political and social changes desired by the people."[58] One major concern of the Court was that imposing strict liability for defamation would chill speech and result in a form of self-censorship.[59] Under such a regime, the press would become hesitant to publish works based on fear of legal ramifications.[60] Even minor, unintentional inaccuracies could deter publication altogether. Because not everything would be shared openly and freely, public debate would also decline, weakening overall democracy; as the Court warned, "fear breeds repression; [ ] repression breeds hate' [ ] hate menaces stable government."[61] The Court therefore explicitly sought to ensure that the press retained the "breathing space" needed "to survive."[62]

This logic applies not only to the institutional press but also to individual speakers. Individuals, like journalists, would become hesitant to criticize public officials freely if libel laws evolved into a tool for punishing political opposition. However, this same "breathing space," when extended to individuals, carries heightened risks in the context of generative AI. Unlike professional journalists, individual speakers are not subject to formalized ethical duties or verification

53. Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) ("[T]he best test of truth is the power of the thought to get itself accepted in the competition of the market . . . .").

54. *See Sullivan*, 376 U.S. at 270 (quoting United States v. Associated Press, 52 F. Supp. 362, 372 (S.D.N.Y. 1943)).

55. *Id.* at 275.

56. *Id.*

57. *See* Stanley v. Georgia, 394 U.S. 557, 564 (1969) ("[The] right to receive information and ideas, regardless of their social worth is fundamental to our free society.") (citation omitted); Whitney v. California, 274 U.S. 357, 377 (1927) ("[T]he remedy to be applied is more speech, not enforced silence.").

58. *Sullivan*, 376 U.S. at 269 (quoting Roth v. United States, 354 U.S. 476, 484 (1957)).

59. *Id.* at 279.

60. *Id.* at 278.

61. *Id.* at 270 (quoting *Whitney*, 274 U.S. at 375–76 (Brandeis J., concurring)).

62. *Id.* at 271–72.

standards as a condition of publishing content online.[63] There is no comparable expectation of due diligence governing individual speech on social media platforms. With advances in generative AI technology, which requires minimal work by the user to provide an immediate output of convincing synthetic media, individuals can disseminate at much greater ease. These realities suggest that courts should reconsider how defamation doctrine, particularly in cases involving public officials, accounts for AI-generated speech when calibrating the scope of constitutional "breathing space."

C. ASSUMPTIONS UNDERLYING CURRENT DEFAMATION LAWS

1. The Ability to Combat

The *Sullivan* framework assumes that public officials possess adequate capacity to respond to and combat libelous statements.[64] By virtue of their status, public officials are presumed to have access to media channels through which they can respond to attacks and correct falsehoods.[65] This assumption justified the Court's decision to heighten the standard for libel claims on the theory that counter-speech, rather than legal recourse, is the appropriate remedy for false speech in the public political sphere.[66] Counter-speech refers to the ability of a speaker to respond through rebuttal, clarification, or correction in the same public forum. Traditionally, public officials have been understood to enjoy larger platforms and greater visibility than private citizens, enabling them to rebut false statements in ways unavailable to most private individuals. This premise, however, is increasingly strained in an AI-driven information environment.

2. The Ability to Retract

The *Sullivan* decision further assumes that false information can be feasibly retracted or corrected through established journalistic norms. This assumption was reasonable in the media environment of *Sullivan*, when public discourse was dominated by professional journalism, particularly newspapers.[67] Professional journalists operate under an ethical obligation to produce accurate information

63. Unlike institutional journalism, social-media platforms primarily rely on reactive, post-publication content moderation systems. *See generally* Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini & Shirin Nilizadeh, *SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice*, BOS. UNIV. OPENBU (2023) (describing platform moderation as ex post "enforcement" of community guidelines).

64. *Sullivan*, 376 U.S. at 270.

65. *Id.* at 304–05 (Goldberg J., concurring).

66. *Id.* at 270, 279 (majority opinion).

67. *See Newspapers Fact Sheet*, PEW RSCH. CTR. (Nov. 10, 2023), https://www.pewresearch.org/journalism/fact-sheet/newspapers (noting that the total estimated circulation of U.S. daily newspapers in 1964 was approximately 60,000,000, as compared to an estimated 20,900,000 in 2022).

and consequently issue public corrections when errors occur.[68] In the contemporary information ecosystem, however, this assumption is increasingly untenable. AI-generated libel spreads rapidly across decentralized social media platforms,[69] often reaching far broader audiences than any subsequent correction or retraction.

3. The Ability to Get the Joke

The Court also rejected strict liability out of concern for overbreadth, fearing that expansive defamation liability would suppress constitutionally protected expression.[70] In particular, the Court emphasized the need to preserve satirical or parodical speech, warning that poorly tailored restrictions risk silencing legitimate political discourse.[71] A key assumption underlying both the *Sullivan* decision and the marketplace of ideas theory is that people are rational actors capable of weighing competing claims.[72] The marketplace of ideas relies on the principle of self-correction, in which, by having free and competitive discourse, errors, biases, and falsehoods will eventually be exposed and weeded out from the truth.[73] At the time of *Sullivan*, journalism frequently employed political cartoons or satirical editorials to critique political figures.[74] The constitutional tolerance for exaggerated or false statements for the sake of First Amendment safeguards rests on the assumption that audiences possess the competence to distinguish literal assertions of fact and satirical or parodical exaggeration.[75] Today, however, generative AI complicates this distinction by producing content that may appear hyper-realistic while remaining ambiguous as to whether it is intended as satire or as an assertion of fact.

a. The Satire Defense

The First Amendment protects satirical expressive content.[76] This protection rests on the assumption that a reasonable viewer would understand

68. *See, e.g.*, *SPJ Code of Ethics*, SOC'Y OF PRO. JOURNALISTS (Sep. 6, 2014, at 16:49 CT), https://www.spj.org/spj-code-of-ethics ("Journalists should [ ] [t]ake responsibility for the accuracy of their work. Verify information before releasing it . . . . Gather, update and correct information throughout the life of a news story.").

69. Social media posts have the potential to go viral in a matter of weeks, days, or even hours depending on audience engagement and algorithmic boosting.

70. *Sullivan*, 376 U.S. at 271–72, 279.

71. *Id.* at 270–72.

72. *See id.* at 270–71; *see also* JOHN STUART MILL, ON LIBERTY 38 (3d ed. 1864) (discussing the human ability to rectify errors through discussion and experience).

73. MILL, *supra* note 72, at 33, 37–38.

74. *See, e.g.*, *Herblock Looks at 1964: Fifty Years Ago in Editorial Cartoons*, LIBR. OF CONG. (last visited Apr. 13, 2026), https://www.loc.gov/exhibits/herblock-gallery/herblock-looks-at-1964.html. For example, the 1964 Presidential Election between Lyndon B. Johnson and Barry Goldwater, as well as the Civil Rights Act of 1964, resulted in political cartoons being extensively used to critique candidates, their platforms, and the legislature.

75. Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 57 (1988).

76. *Id.* at 53–54.

that a satirical publication is not an assertion of fact.[77] To assess this, courts utilize the reasonable person standard.[78] Satire that is so exaggerated as to preclude belief is presumed incapable of causing reputational harm.[79] When evaluating satire in defamation cases, courts must consider "the right of individuals and groups to be free from attack on their property, dignity, and honor versus the right of individuals to free expression."[80] In most cases, utilizing the reasonable person standard, parody and satire are treated as opinions and therefore cannot be defamatory.

For example, in California's Civil Jury Instructions Section 1707, "Fact Versus Opinion," the jury is guided to "consider whether the *average* [reader/listener] would conclude from the language of the statement and its context that [the defendant] was implying that a false statement of fact is true."[81] This instruction is only given to the jury "if the court concludes that a statement could reasonably be construed as implying a false assertion of fact."[82] The instructions further elaborate that "rhetorical hyperbole," "vigorous epithet[s]," "lusty and imaginative expression[s] of contempt," and "language used 'in a loose, figurative sense'" have all been granted constitutional protection.[83] AI-generated libel, however, strains the workability of the reasonable person standard by producing content that appears realistic rather than exaggerated and thereby blurring the line between satire and factual assertion.

II. THE RIDE OF AI LIBEL

Generative AI presents a profound challenge to First Amendment jurisprudence regarding defamatory speech against political figures by testing the workability of the actual malice standard. As social media has become a dominant source of news, the origins of potential libel suits no longer arise solely from traditional media outlets. Millions of individuals who use social media daily now have quick access to generative AI tools, enabling them to create and disseminate synthetic content instantly through their personal accounts.

Walters v. OpenAI[84] illustrates both the legal complexities of AI-generated defamation and the practical difficulty of succeeding in such suits, with the court ultimately dismissing the claim for failure to plausibly allege falsity, fault, and

77. *Id.* at 57.

78. Catherine L. Amspacher & Randel Steven Springer, Note, *Humor, Defamation and Intentional Infliction of Emotional Distress: The Potential Predicament for Private Figure Plaintiffs*, 31 WM. & MARY L. REV. 701, 727 (1990); *see also Hustler Mag., Inc.*, 485 U.S. at 57 (accepting a jury's finding that an ad parody was not reasonably believable).

79. *Hustler Mag., Inc.*, 485 U.S. at 54, 57.

80. Laura E. Little, *Just a Joke: Defamatory Humor and Incongruity's Promise*, 21 S. CAL. INTERDISC. L.J. 95, 95 (2011).

81. Jud. Council of Cal. Civ. Jury Instructions, CACI No. 1707, *Fact Versus Opinion* (2025) (emphasis added).

82. *Id.*

83. *Id.* at 1057.

84. *See* Walters v. OpenAI, L.L.C., No. 23-13843, 2004 WL 1366490, at *1 (11th Cir. 2024).

a cognizable theory of liability under existing defamation doctrine.[85] *Walters* highlights unresolved questions about how liability should be allocated and whether existing doctrinal frameworks are adequate to address AI-generated speech. In today's information environment, the foundational assumptions that shaped *Sullivan*—the effectiveness of counter speech, the feasibility of retraction, and the public's ability to distinguish satire from reality—are increasingly undermined. Accordingly, the rise of AI-generated libel warrants a renewed examination of defamation law, particularly as deepfake technology erodes traditional legal distinctions among fact, opinion, and satire.

A. GENERATIVE AI IN TODAY'S POLITICAL ARENA

Recent studies reveal that 53 percent of U.S. adults obtain news, at least occasionally, from social media, with Facebook and YouTube serving as the two most common platforms for regular news consumption.[86] As a result, political information—accurate or otherwise—is increasingly broadcasted through algorithmic platforms rather than traditional journalistic institutions. Many people who engage with political news, and even those who encounter political content incidentally on social media, are likely to be familiar with the growing use of deepfakes involving public officials.[87]

Recent incidents illustrate the pervasiveness of AI-generated political media. In one example, an AI-generated video of President Trump kissing Elon Musk's feet, accompanied by the heading "Long Live the Real King," circulated throughout the Department of Housing and Urban Development headquarters in February 2025.[88] Similarly, in January 2024, generative AI infiltrated the New Hampshire presidential primaries when a robocall mimicking the voice of then-President Joe Biden urged residents not to vote in the presidential primary, falsely suggesting that participation should be reserved for the general election.[89] This incident was widely characterized as an attempt to suppress voters from putting then-President Biden's name on their presidential primary ballot.[90] However, the use of deepfakes is not limited to the 2024 election cycle.

---

85. *See* Order Granting Summary Judgment in Favor of Defendant OpenAI, L.L.C. at 5–18, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. May 19, 2025).

86. PEW RSCH. CTR., *supra* note 26.

87. *See, e.g.*, Tvesha Sippy, Florence E. Enock, Jonathan Bright & Helen Z. Margetts, *Behind the Deepfake: 8% Create; 90% Concerned*, ALAN TURING INST., June 2024, at 7–9 (noting survey respondents encountered deepfakes on social media or expressed concern about their potential misuse).

88. Molly Bohannon, *AI-Generated Video of Trump Kissing Elon Musk's Feet Reportedly Plays in HUD Building*, FORBES (Feb. 24, 2025, at 18:16 ET), https://www.forbes.com/sites/mollybohannon/2025/02/24/video-of-trump-kissing-elon-musks-feet-reportedly-plays-in-hud-building.

89. Kevin Collier & Scott Wong, *Fake Biden Robocall Telling Democrats Not to Vote Is Likely an AI-Generated Deepfake*, NBC NEWS (Jan. 22, 2024, at 16:12 PT), https://www.nbcnews.com/tech/misinformation/joe-biden-new-hampshire-robocall-fake-voice-deep-ai-primary-rcna135120.

90. *Id.* In the call, a voice similar to that of then-President Joe Biden, advised listeners that "it's important that you save your vote for the November election . . . your vote makes a difference in November, not this Tuesday." Alex Seitz-Wald & Mike Memoli, *Fake Joe Biden Robocall Tells New Hampshire Democrats Not to*

In 2023, Florida Governor Ron DeSantis circulated three AI-generated images depicting President Trump and then-lead member of the White House Coronavirus Task Force Dr. Anthony Fauci kissing.[91] More broadly, generative AI tools are becoming integrated into everyday digital uses, including search engines such as Microsoft's Bing and Google's Bard[92], making exposure to synthetic political content increasingly unavoidable.

Although political figures accept heightened scrutiny by assuming positions of public power[93], AI-generated libel has emerged as a quick, low-cost mechanism for inflicting reputational harm. Such content also harms public viewers who fall victim to believing disinformation. This distorts political understanding and undermines trust in democratic processes.[94] The spread of deepfake technology is further facilitated by social media's scale and virality. As deepfakes increasingly blur the distinction between authentic and fabricated content, these developments compel a reassessment of the assumptions underlying the Supreme Court's decision in *Sullivan* and whether those assumptions remain viable today.

B. *WALTERS V. OPENAI, L.L.C.*—THE FIRST OF MANY?

The first defamation lawsuit brought against an AI generator occurred in *Walters*.[95] The case centers on allegedly hallucinated[96] defamatory outputs about Mark Walters, a radio talk show host.[97] Walters alleged that ChatGPT

Vote Tuesday, NBC NEWS (Jan. 22, 2024, at 08:45 PT), https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984.

91. Nicholas Nehamas, *DeSantis Campaign Uses Apparently Fake Images to Attack Trump on Twitter*, N.Y. TIMES (June 8, 2023), https://www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html.

92. Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 493, 498 (2023); Christina LaChapelle & Catherine Tucker, *Generative AI in Political Advertising*, BRENNAN CTR. FOR JUST. (Nov. 28, 2023), https://www.brennancenter.org/our-work/research-reports/generative-ai-political-advertising.

93. *See* New York Times Co. v. Sullivan, 376 U.S. 255, 270 (1964); *see also* Gertz v. Welch, 418 U.S. 323, 344 (1974) ("An individual who decides to seek governmental office . . . runs the risk of closer public scrutiny than might otherwise be the case.").

94. *See generally* Maria Pawelec, *Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions*, DIGIT. SOC'Y, Sep. 2022, at 1 (discussing how deepfakes weaken democratic functions and hamper efforts to hold political officials accountable); *see also* Dan Cavedon-Taylor, *Deepfakes and Democracy: A Catch-22?*, 11 J. AM. PHIL. ASS'N 447, 447–48 (2025) (analyzing the deleterious effect that deepfakes have on democratic elections from a nihilistic perspective).

95. *Gibson Dunn Wins Significant Victory for Client OpenAI Defending Against Defamation Claim Based on "Hallucinated" Generative AI Output*, GIBSON DUNN (May 21, 2025), https://www.gibsondunn.com/gibson-dunn-wins-significant-victory-for-client-openai-defending-against-defamation-claim-based-on-hallucinated-generative-ai-output.

96. Hallucination refers to generative AI's ability to "make up" outputs, ultimately presenting false information as fact. AI may hallucinate due to insufficient data or training of the large language model, or lack of guidance from the prompter. *See When AI Gets It Wrong: Addressing AI Hallucinations and Bias*, MIT MGMT. STS TEACHING & LEARNING TECHS., https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias (last visited Apr. 13, 2026); *see also* Leslie Y. Garfield Tenzer, *Defamation in the Age of Artificial Intelligence*, 80 NYU ANN. SURV. AM. L. 135, 138 (2024) (highlighting the difficulty plaintiffs may face in identifying defendants and proving the elements of a defamation claim when AI sources produce defamatory hallucinations).

97. Complaint at 4, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. June 5, 2023).

generated defamatory content about him after being prompted by a ChatGPT user.[98] According to Walters, the entirety of the statements provided in ChatGPT's summary were false.[99]

Fred Riehl, a journalist, approached ChatGPT to obtain more information about a lawsuit he was reporting on.[100] The lawsuit in question, *The Second Amendment Foundation v. Robert Ferguson*, was filed in a federal court in the Western District of Washington.[101] Walters was not a party to this suit.[102] Riehl submitted a URL linking to the complaint from the Second Amendment Foundation's website and then asked ChatGPT to summarize the allegations contained in the filing.[103] In response, ChatGPT incorrectly summarized the document as a legal complaint against Mark Walters, stating that Walters was "accused of defrauding and embezzling funds from the [Second Amendment Foundation]."[104]

When Riehl then prompted ChatGPT to provide the portion of the complaint relating to Walters, ChatGPT proceeded to hallucinate an entire paragraph from the complaint. Its response asserted that:

> Walters has served as the Treasurer and Chief Financial Officer of SAF since at least 2012. Walters has access to SAF's bank accounts and financial records and is responsible for maintaining those records and providing financial reports to SAF's board of directors. Walters owes SAF a fiduciary duty of loyalty and care, and is required to act in good faith and with the best interests of SAF in mind. Walters has breached these duties and responsibilities by, among other things, embezzling and misappropriating SAF's funds and assets for his own benefit, and manipulating SAF's financial records and bank statements to conceal his activities.[105]

None of these statements appeared in the actual complaint, and the hallucinated passage even included an incorrect case number.[106] Walters sought monetary damages from OpenAI for the reputational harm that was allegedly caused by ChatGPT's fabricated output.[107] In November 2024, OpenAI filed a motion for summary judgment;[108] in May 2025, this motion was granted in favor of OpenAI.

This case illustrates the growing legal complexities facing public figures seeking redress for AI-generated defamation. It raises fundamental questions

98. *Id.* at 2–3.

99. *Id.* at 3.

100. Plaintiff's Response in Opposition to Defendant's Motion to Dismiss Amended Complaint at 5, Walters v. OpenAI, L.L.C., No. 23-CV-03122 (N.D. Ga. Oct. 16, 2023).

101. *Id.* at 6.

102. *Id.*

103. *Id.* at 6–7.

104. *Id.* at 6.

105. *Id.* at 8–9.

106. *Id.* at 9.

107. Complaint at 4, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. June 5, 2023).

108. Defendant OpenAI, L.L.C.'s Memorandum of Law in Support of Motion for Summary Judgment at 26, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. Nov. 14, 2024).

about attribution and liability: Who, if anyone, should be held responsible for hallucinated defamation? The AI developer, those who create the underlying algorithm, or the end user who prompted the system? The AI user may also act recklessly by relying on unverified AI outputs and further compounding that recklessness by disseminating those outputs to the public. In *Walters*, the ChatGPT user had no intention to produce defamatory content, and it has been widely argued that holding AI developers liable for false outputs presents substantial doctrinal and practical obstacles.[109] The analysis becomes more complex, however, when an AI user *does* have this intention and deliberately prompts a system to produce content that is known to be false. Actual malice requires that the AI user either knows its output is false or acts with reckless disregard for its truth.[110] When a user knowingly supplies false premises or intentionally prompts AI to generate fabricated media presented as real, there is convincing evidence of actual malice that would likely satisfy the *Sullivan* standard.

C. ASSUMPTIONS UNDERLYING TODAY'S REALITY

The dominant form of libel at the time of *Sullivan* consisted primarily of written, text-based defamation published in newspapers.[111] Although the Court has addressed defamation and reputational harm arising from parody and other nontraditional forms of expression, it has never confronted cases involving falsified images, audio, or video that convincingly mimic reality. Nor has the Court addressed defamation arising from AI-altered or entirely AI-generated media.

The presiding Court in *Sullivan* could not have anticipated the rapid technological developments that now allow deepfakes to challenge traditional defamation doctrine. Now that synthetic media is a defining feature of today's information ecosystem, it is critical for courts and legislatures to reassess which assumptions underlying modern defamation law remain valid, which require modification, and which no longer apply. The rapid evolution of AI technology has drastically altered the context in which speech is circulated to the public. Although *Sullivan*'s core commitment to protecting political speech remains

109. *See* Joel Simon, *Can AI Be Sued for Defamation?*, COLUM. JOURNALISM REV. (Mar. 18, 2024), https://www.cjr.org/analysis/ai-sued-suit-defamation-libel-chatgpt-google-volokh.php; *see also* Grayson Horton, *AI and Defamation: Who Do You Sue?*, AVE MARIA SCH. OF L. (May 17, 2024), https://www.avemaria law.edu/who-do-you-sue (discussing *Walters* and the obstacles plaintiffs will face, such as identification, whether the statement was one of fact, and the calculation of damages). *But see* Garfield Tenzer, *supra* note 96, at 169 (arguing that the principal of the generative AI company or the employer, on behalf of the owner, can be subject to liability).

110. *See* N.Y. Times Co. v. Sullivan, 376 U.S. 254, 280 (1964) ("'[A]ctual malice' [means] with knowledge that it was false or with reckless disregard of whether it was false or not.").

111. *See* Samantha Barbas, *The Press and Libel Before* New York Times v. Sullivan, 44 COLUM. J.L. & ARTS 511, 534–38 (2021) (discussing the resurgence of newspapers in coverage of local news and the increase in libel cases against the press).

central to First Amendment jurisprudence, new forms of harm now call into question the continued viability of the assumptions on which the doctrine rests.

1. The Ability to Combat and Retract, Undermined by Virality

In the mid-twentieth century, the media landscape and news dissemination were concentrated in a small number of large institutions that controlled information flows and set collective agendas.[112] Today, the rise of social media and the increasingly global interconnectedness have, in some respects, democratized speech by lowering barriers to participation, expanding access to information, and enabling broader engagement in political discourse.[113] Despite these advances, it has never been easier to disseminate false information to a mass audience; digital platforms and their algorithms reward engagement and virality, which can accelerate the spread of misinformation and distort public understanding.[114] Narrative control has shifted from editorial judgment to algorithmic prioritization. A single viral defamatory image or post can dismantle a reputation, sometimes permanently, particularly for political public figures. The scale of information spread—unimaginable at the time of *Sullivan*—makes countering viral disinformation more difficult, and when falsehoods reach massive audiences through social amplification, counter-speech rarely achieves comparable reach.

The internet is functionally permanent. Once posted, AI-generated libel is often unretractable, as false or misleading content continues to circulate even after corrections or removals are issued.[115] Screenshots, retweets, reposts, and digital archives preserve false statements long after their original publication. Many times, this content may be missing crucial context. The assumption that false speech can be promptly corrected or retracted no longer carries the same weight. Moreover, because algorithmic systems are designed to maximize engagement rather than accuracy,[116] false or sensational content is often amplified more forcefully than subsequent corrections. These systems reward novelty, outrage, and entertainment—conditions that favor misinformation over truth.[117] Finally, algorithmic systems reinforce disparities in online visibility by

112. *See* SAMANTHA BARBAS, ACTUAL MALICE: CIVIL RIGHTS AND FREEDOM OF THE PRESS IN *NEW YORK TIMES V. SULLIVAN* 7 (2023) (noting the breadth and reach of the New York Times).

113. *See* Sneha Gubbala & Sarah Austin, *Majorities in Most Countries Surveyed Say Social Media Is Good for Democracy*, PEW RSCH. CTR. (Feb. 23, 2024), https://www.pewresearch.org/short-reads/2024/02/23/majorities-in-most-countries-surveyed-say-social-media-is-good-for-democracy.

114. Shin & Shin, *supra* note 11, at 1441 (noting how virality functions as a mechanism of misinformation, "transforming isolated claims into widespread societal narratives").

115. *See, e.g.*, Jiexun Li & Xiaohui Chang, *Combating Misinformation by Sharing the Truth: A Study on the Spread of Fact-Checks on Social Media*, 25 INFO. SYS. FRONTIERS 1479, 1480 (2023) (finding that corrective messages often lag behind misinformation spread on social media).

116. *See* Maria Alessandra Golino, *Algorithms in Social Media Platforms*, INST. FOR INTERNET & THE JUST SOC'Y (Apr. 24, 2021), https://www.internetjustsociety.org/algorithms-in-social-media-platforms.

117. *See* Killian L. McLoughlin, William J. Brady, Aden Goolsbee, Ben Kaiser, Kate Klonick & M. J. Crockett, *Misinformation Exploits Outrage to Spread Online*, 386 SCI. 991, 995 (2024) ("Because outrage is

prioritizing content that generates high engagement. Users with established audiences are more likely to receive algorithmic amplification, further expanding their reach while making subsequent corrective speech substantially less likely to achieve comparable exposure.[118] Although false claims may eventually be challenged or corrected, those corrections struggle to achieve reach or engagement equivalent to that of the original content. As a result, the premise that the marketplace of ideas will naturally self-correct is significantly weakened, particularly as deepfakes increasingly obscure the boundary between truth and falsity.

2. Satire and the Collapse of the Reasonable Person Standard

In response to the plaintiff's claim in *Walters*, OpenAI argued that no reasonable user of ChatGPT would believe its outputs to be factual statements.[119] This contention, however, does not reflect how users actually interact with generative AI systems. In its amicus brief, NYU's Technology Law & Policy Clinic asserted that a reasonable user of generative AI systems could, in fact, believe AI outputs to be accurate given the increasing trust between the public and generative AI systems.[120] OpenAI provides a disclaimer in its Terms of Use that ChatGPT may "result in [o]utput that does not accurately reflect real people, places, or facts,"[121] but such disclaimers do little to actually stop reasonable users from believing the accuracy of the outputs.[122]

Moreover, courts evaluating defamation claims focus on whether a reasonable audience would understand the challenged expression as stating facts rather than opinion, exaggeration, or satire. As courts have explained, the "key inquiry is whether the challenged expression . . . would reasonably appear to state or imply assertions of objective fact."[123] AI-generated media confounds this analysis because it exploits fundamental neurological and psychological differences in how humans process words versus images, audio, and video. Human cognition processes visual and auditory media differently, and often

associated with increased engagement online, outrage-evoking misinformation may be likely to spread farther in part because of the algorithmic amplification of engaging content.").

118. *See id.*; *see also* FORBES AGENCY COUNCIL, *The Algorithm: Understand It, Leverage It* (Nov. 6, 2024), https://councils.forbes.com/blog/understanding-and-leveraging-the-algorithm (describing how platform algorithms consider follower count in determining reach, prioritizing engagement signals such as likes, comments, and shares).

119. Defendant's Motion for Summary Judgment at 12, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. Nov. 14, 2024).

120. Brief for Tech. L. & Pol'y Clinic at NYU Sch. of L. as Amicus Curiae Supporting Neither Party at 2, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. Nov. 14, 2024).

121. *Terms of Use*, OPENAI, https://openai.com/policies/terms-of-use/ (last visited Apr. 13, 2026).

122. *See* Peter Henderson, Tatsunori Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 638 (2023) (noting that many users will likely dismiss the small, one-lined, buried disclaimers as "boilerplate").

123. Takieh v. O'Meara, 497 P.3d 1000, 1006 (Ariz. Ct. App. 2021) (emphasis removed) (quoting Yetman v. English, 811 P.2d 323, 331 (Ariz. 1991)).

more powerfully, than text.[124] As the adage "a picture is worth a thousand words"[125] suggests, images exert a unique persuasive force. Visual and auditory stimuli are perceived more immediately, are often emotionally charged, and tend to be more memorable.[126] These qualities can imbue them with a heightened sense of credibility.

Scientific research has long established that the human ability to remember pictures is much stronger than the human ability to remember words.[127] A seminal 1970 study demonstrated that individuals could recall more than 2,000 pictures with at least 90 percent accuracy, even after several days had elapsed between learning and testing.[128] Participants viewed 2,560 photographic stimuli for just ten seconds each.[129] Results showed that even when the presentation time was reduced to one second per picture, performance was not seriously affected.[130] Subsequent research has reaffirmed this result. A 1998 study revealed that people remember pictures better than words because words, unlike images, activate left hemisphere regions of the brain known for language tasks, increasing the difficulty of later retrieval from episodic memory.[131] Some researchers theorize that this superiority in memory derives from the fact that pictures trigger "multiple representations and associations with other knowledge about the world," consequently requiring "more elaborate encoding" than words.[132] More recently, a 2014 study conducted by the Massachusetts Institute of Technology found that individuals could detect and retain information about images shown for only thirteen milliseconds.[133]

Modern information consumption further amplifies these cognitive differences. In an attention economy optimized for speed, users frequently encounter images and short videos without engaging with accompanying text.[134] Why read an entire article when a single picture on the front page can convey its essential message? Given the speed at which neural processing occurs, even

124. *See* ALLAN PAIVIO, IMAGERY AND VERBAL PROCESSES 391 (1971).

125. This phrase has been attributed to Frederick R. Barnard, who first titled a 1921 advertisement "One Look is Worth a Thousand Words," later adapting it in a 1927 advertisement to "One Picture is Worth Ten Thousand Words." *See* Daryl H. Hepting, *A Picture's Worth*, UNIV. OF REGINA, https://www2.cs.uregina.ca/~hepting/projects/pictures-worth (last visited Apr. 13, 2026).

126. PAIVIO, *supra* note 124, at 387.

127. *Id.*

128. Lionel Standing, Jerry Conezio & Ralph Norman Haber, *Perception and Memory for Pictures: Single-Trial Learning of 2500 Visual Stimuli*, 19 PSYCHONOMIC SCI. 73, 73–74 (1970).

129. *Id.* at 73.

130. *Id.*

131. Cheryl L. Grady, Anthony R. McIntosh, M. Natasha Rajah & Fergus I. M. Craik, *Neural Correlates of the Episodic Encoding of Pictures and Words*, 95 PROC. NAT'L ACAD. SCI. USA 2703, 2706–07 (1998).

132. *Id.* at 2703.

133. Mary C. Potter, Brad Wyble, Carl Erick Hagmann & Emily S. McCourt, *Detecting Meaning in RSVP at 13 Ms per Picture*, 76 ATTENTION PERCEPTION & PSYCHOPHYSICS 270, 277 (2014).

134. *See, e.g.*, Xinran Dai & Jing Wang, *Effect of Online Video Infotainment on Audience Attention*, HUMANS. & SOC. SCIS. COMMC'NS, July 17, 2023, at 2; Xinzhou Xie, Yanjun Lin & Qiyu Bai, *Ephemeral Emotional Resonance: User-Perceived Functional Value Leading to Short-Form Video Use*, BEHAV. SCIS., Mar. 10, 2025, at 10.

those who skip textual explanations are still highly likely to remember the image on the front page.[135] Importantly, image interpretation is not uniform. Individual perception through neural processing is influenced by past subjective experiences, emotions, understandings, or meanings invoked.[136] Visual perception engages not only the brain's visual cortex but also regions responsible for emotional processing, including the amygdala, enabling images to influence viewers on cognitive and emotional levels simultaneously[137] The "Picture Superiority Effect" explains the phenomenon that dual-coding of the brain channels—combining visual and semantic processing—enhances memory retention of information presented in images more than information presented in words.[138] Further, perceptually rich images enjoy an additional advantage due to their distinctiveness.[139]

These cognitive dynamics significantly intensify the harm posed by AI-generated libel. Consider the following scenario: You read a news article alleging that a public official was caught snorting cocaine in a bar over the weekend. This may prompt skepticism or disbelief; you may cautiously believe the story, or you may outright scoff. Now consider the same allegation, accompanied by a photograph of the public official, leaning down on the bar counter, their nose millimeters away from three distinct white lines of powder. Now, you might feel different about the news. The visual "evidence" may trigger immediate belief and moral condemnation. Public outrage follows. What viewers do not realize is that the alleged proof is actually a 100 percent doctored image created with deep-fake technology. But it looks *so* real. And even after the falsity of the image's origin is revealed, the reputational damage to the public figure has already been done. The image has circulated widely, imprinted itself in public memory, and reshaped perception in ways that cannot be undone through later correction.

a. The Satire Defense

The psychological differences between the perception of words and images raise substantial concerns about the continued viability of the satire defense in response to defamation claims. Not only do images exert a demonstrably different and more powerful cognitive impact than text, but advancements in deepfake technology amplify the potential for reputational harm at an unprecedented scale. With the rise of deepfake technology, the assumption that

135. Potter et. al., *supra* note 133, at 276.

136. Margarida Alpuim & Katja Ehrenberg, *Why Images Are So Powerful—and What Matters When Choosing Them*, BONN INST. (Aug. 3, 2023), https://www.bonn-institute.org/en/news/psychology-in-journalism-5.

137. *Id.*

138. Allan Paivio & Kalman Csapo, *Picture Superiority in Free Recall: Imagery or Dual Coding?*, 5 COGNITIVE PSYCH. 176, 200 (1973).

139. Joyce M. Oates & Lynne M. Reder, *Memory for Pictures: Sometimes a Picture Is Not Worth a Single Word*, *in* SUCCESSFUL REMEMBERING AND SUCCESSFUL FORGETTING: A FESTSCHRIFT IN HONOR OF ROBERT A. BJORK 447, 449 (Aaron S. Benjamin ed., 2010).

a reasonable viewer can reliably discern truth from fiction is increasingly undermined. A fabricated image or video may be obviously parodical entertainment in some contexts, yet in others may appear so realistic that, because of the psychological impact that imagery has on the viewer, it can deceive viewers into accepting it as a literal fact. Revisiting the reasonable person standard in this context reveals tension. The wide range of psychological interpretations of visual content suggests that it may no longer be reasonable to apply a single, blanket interpretive framework when assessing AI-generated content.

The Supreme Court has repeatedly recognized that an informed electorate is essential to the functioning of a democratic system.[140] The Court has upheld disclosure and transparency requirements in the political sphere, including laws governing campaign finance,[141] political advertising,[142] and ballot initiative disclosures.[143] Deepfake technology threatens these democratic interests by corrupting the available information on which voters can rely. Because AI-generated images, audio, and video can be highly persuasive, placing responsibility solely on the viewers and voters to verify authenticity imposes an increasingly unrealistic burden. False depictions of candidates may mislead voters about a candidate's character, conduct, or policy positions. Over time, these dynamics risks eroding public trust in legitimate media institutions and may also lead audiences to dismiss accurate reporting as fabricated, fostering cynicism rather than informed skepticism.[144] In this way, the misuse of AI-generated satire not only challenges existing defamation doctrine but also threatens the informational foundations upon which democratic self-governance depends.

III. REFORMS

The assumptions that have long underpinned modern defamation doctrine—effective counter-speech, feasible retraction, and a rational audience capable of distinguishing satire from fact—are increasingly ill-suited to an information environment shaped by generative AI. As such, it is important to look at the current attempts being made at the state level to address AI-generated libel and to evaluate whether these emerging regulatory responses meaningfully respond to the doctrinal gaps identified above.

---

140. *See* Buckley v. Valeo, 424 U.S. 1, 14–15 (1976) ("In a republic where the people are sovereign, the ability of the citizenry to make informed choices among candidates for office is essential, for the identities of those who are elected will inevitably shape the course that we follow as a nation.").

141. *See id.* at 82.

142. *See* Citizens United v. FEC, 558 U.S. 310, 371 (2010).

143. *See* Doe v. Reed, 561 U.S. 186, 191 (2010).

144. *See, e.g.*, Emma Hoes, Bernhard Clemm, Theresa Gessler, Sijia Qian & Magdalena Wojcieszak, *(Media Attention to) Misinformation Can Undermine Trust in Scientists*, POL. BEHAV., Nov. 10, 2025, at 23 (noting that exposure to misinformation and news coverage about misinformation can lower trust in credible sources of knowledge, including even reputable institutions and scientists).

A. CURRENT ATTEMPTS AT CONTAINING AI-LIBEL

To date, most legislative responses to AI-generated misinformation and defamation have emerged at the state level, focusing primarily on election-related uses of generative AI through mandatory disclosure regimes.[145] Mandatory disclosure laws require entities or individuals to reveal explicit information to increase transparency and accountability.[146] As of December 2025, more than twenty states have enacted some form of legislation for AI-related mandatory disclosure.[147] For example, California Government Code Section 84514, enacted in 2024, regulates political advertisements on digital platforms.[148] This Section requires committees to add a disclosure to political advertisements created using AI: "Ad generated or substantially altered using artificial intelligence."[149] Similarly, New York's Election Law Section 14-106 requires political communications that contain materially deceptive media, including AI-generated content, to disclose that the media has been manipulated.[150] Legislative disclosure requirements are not the only mechanisms proposed to curb AI-generated misinformation and defamation. Private actors, particularly technology companies, have also explored the use of AI as a content moderation tool.[151] This includes using AI-driven algorithms to analyze large volumes of online content and detect potentially false, misleading, or manipulated material.[152] Although imperfect, such tools may assist platforms in flagging or slowing the dissemination of harmful content before it achieves viral reach.

Despite their appeal, these approaches largely sidestep the core doctrinal problem identified in Part II: They regulate disclosure and dissemination without addressing how defamation laws should treat AI-generated falsities when reputational harm has already occurred. Mandatory disclosures and platform-level safeguards, while valuable, are insufficient substitutes for a coherent recalibration of defamation doctrine itself.

---

145. *See Deepfakes in Elections and Campaigns*, NAT'L CONF. OF STATE LEG. (Dec. 15, 2025), https://www.ncsl.org/elections-and-campaigns/artificial-intelligence-ai-in-elections-and-campaigns.

146. *See* VALERIE C. BRANNON, VICTORIA L. KILLION, WHITNEY K. NOVAK & L. PAIGE WHITAKER, CONG. RSCH. SERV., IF12388, FIRST AMENDMENT LIMITATIONS ON DISCLOSURE REQUIREMENTS (Apr. 26, 2023).

147. NATL. CONF. OF STATE LEG., *supra* note 145.

148. CAL. GOV. CODE § 84514 (West 2025).

149. *Id.* § 84514(a)(1).

150. N.Y. ELEC. LAW § 14-106 (McKinney 2026).

151. *See* Sarah A. Fisher, Jeffrey W. Howard & Beatriz Kira, *Moderating Synthetic Content: The Challenge of Generative AI*, 37 PHIL. & TECH. 133 (2024) (discussing AI-assisted content moderation by online platforms as a means of addressing synthetic content).

152. *See, e.g.*, *id.*; National Institute of Standards and Technology, *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency* 6–11 (2024) (describing technical methods for detecting, labeling, authenticating, and mitigating synthetic media).

B. WHY EXISTING REFORMS FAIL IN AN INTERNET-DRIVEN ECOSYSTEM

Although recent state legislation has started to address the use of AI-generated media in election advertisements, these measures leave largely untouched the casual yet pervasive use of generative AI on social media platforms, which increasingly function as de facto news outlets.[153] It may be feasible for legislation to enforce mandatory disclosures against regulatory bodies or public entities, but attempts to regulate actions on a private, individual level are more complex due to the level of constitutional scrutiny applied to such laws.[154] As a result, existing reforms capture only a narrow slice of AI-generated disinformation while leaving the most common vectors of AI-libel unregulated.

1. Constitutional Limits on Regulating Private AI Use

Strict scrutiny[155] applies to legislation that attempts to regulate speech on the basis of its content or message,[156] as well as to laws that compel speech where the speaker would otherwise remain silent.[157] Mandatory disclosure requirements obligating private individuals to label AI-generated images in personal, noncommercial contexts are therefore likely to trigger strict scrutiny review.[158] This is particularly true when the compelled disclosure burdens artistic, political, or other expressive activity.[159] Under this standard, the government must justify the disclosure with a compelling government interest,[160] such as ensuring transparency and protecting the integrity of the electoral process.[161] Although these interests are significant, strict scrutiny remains an exacting standard. A universal disclosure mandate covering all AI-generated content, imposed on all individuals regardless of context, is unlikely to be deemed narrowly tailored. To survive constitutional review in federal courts, such a regulation would need to be carefully limited to harmful AI-generated content and structured to avoid unduly burdening protected speech, including satire, parody, and core political expression.

153. For a compilation of recent state laws, see NAT'L CONF. OF STATE LEG., *supra* note 145.

154. Compelled disclosures that require individuals to speak in a particular way are treated as content-based speech regulations, triggering strict scrutiny. *See* Nat'l Inst. of Fam. & Life Advocs. v. Becerra, 585 U.S. 755, 766 (2018).

155. Strict scrutiny is the most demanding level of review a court uses to evaluate a law. This standard "requires the Government to prove that the restriction furthers a compelling interest and is narrowly tailored to achieve that interest." Reed v. Town of Gilbert, 576 U.S. 155, 171 (2015) (quoting Ariz. Free Enter. Club's Freedom Club PAC v. Bennett, 564 U.S. 721, 734 (2011)). Most laws reviewed under strict scrutiny do not survive.

156. *Id.* at 163.

157. *See, e.g.*, *Nat'l Inst. of Fam. & Life Advocs.*, 585 U.S. at 766; Wooley v. Maynard, 430 U.S. 705, 714 (1977).

158. *See Nat'l Inst. of Fam. & Life Advocs.*, 585 U.S. at 766.

159. *See id.*

160. *See* United States v. Playboy Ent. Grp., Inc., 529 U.S. 803, 813 (2000).

161. *See* Doe v. Reed, 561 U.S. 186, 199 (2010); McCutcheon v. FEC, 572 U.S. 185, 206–09 (2014).

2. The Attribution Problem: Difficulty in Identifying the Speaker

One of the most significant obstacles in AI-libel claims is identifying the responsible speaker. Social media anonymity, combined with the often-attenuated relationship between the original AI-user and subsequent disseminators, complicates attribution in ways not contemplated by traditional defamation doctrine. In some cases, such as Florida Governor DeSantis's circulation of AI-generated images depicting President Trump and Dr. Fauci, the point of origin is readily identifiable.[162] However, attribution is far more opaque in other contexts, such as the recent incident where government computers were hacked and used to display AI-generated videos of President Trump and Elon Musk.[163] Many social media platforms permit users to remain anonymous or pseudonymous[164], undermining the ability to trace defamatory content back to its originator and hold that actor accountable. This attribution problem is compounded when the publisher utilizes a generative AI system as an intermediary. Rather than authoring defamatory content directly, the user delegates its creation to an AI model. This inserts an additional layer between the human decision-maker and the final output. Although this delegation complicates evidentiary attribution, it should not shield the user from liability. Generating defamatory media through an AI platform, rather than manually, does not negate authorship or reduce culpability under defamation principles; defamation law focuses not on *how* the words are produced, but rather *who* published them and with *what* level of fault.[165]

3. Liability Fractures Across Platforms, Users and Algorithms

Attribution challenges are made more complex by the mechanics of dissemination on social media platforms. A single instance of AI-generated defamation may involve multiple actors—an AI developer, the prompting user, and countless downstream amplifiers who contribute to the spread through reposts, shares, and algorithmic engagement. Under current law, Section 230 of the Communications Decency Act of 1996 largely shields online platforms from liability for third-party content.[166] Through this Act, Congress intended to promote the rapid growth of the internet, while encouraging voluntary content moderation without treating platforms as publishers of user-generated speech.[167]

---

162. *See* Nehamas, *supra* note 91.

163. Bohannon, *supra* note 88.

164. *See, e.g.*, Reddit Help, *How Does Being Anonymous Work on Reddit?*, https://support.reddithelp.com/hc/en-us/articles/7420342178324-How-does-being-anonymous-work-on-Reddit (last visited Aug. 3, 2026) (explaining that the vast majority of Reddit users choose usernames that do not reveal their identities); X Help Center, *How to Change Your X Username*, https://help.x.com/en/managing-your-account/change-x-handle (last visited Aug. 3, 2026) (explaining that users select unique public usernames and separate display names, which may—but need not—be a real name).

165. *See* Gregory M. Dickinson, *Law Proofing the Future*, 63 HARV. J. ON LEGIS. 81, 100 (2026).

166. 47 U.S.C. § 230(c).

167. *See id.* §§ 230(b), (c)(1).

This makes it nearly impossible to impose a general duty of responsibility on platforms to regulate defamatory AI-generated content. Liability may attach, however, where a platform materially contributes to or affirmatively encourages defamatory use.[168]

A related consideration is the increasing use of generative AI by public officials themselves. Public officials occupy positions of public trust and are traditionally held to heightened standards of candor and responsibility when communicating in their official capacities.[169] This should be particularly pertinent where their statements have the potential to influence public understanding of political or governmental affairs. Yet more frequently, we see the White House and the President casually disseminating AI-generated images to official pages on media platforms such as Instagram.[170] In this context, it would be inconsistent and normatively troubling to impose heightened responsibility on private individuals for AI-generated content while exempting public officials—whose speech carries institutional authority and greater potential for harm—from comparable accountability.

C. PRESERVING THE MARKETPLACE OF IDEAS IN AN AI-DISTORTED ENVIRONMENT

Who bears the burden of verifying the truth of information introduced into the modern marketplace of ideas? The marketplace of ideas traditionally assumes that even distasteful or false ideas must be permitted to enter public discourse to compete with truthful ones, placing the burden of discernment on listeners rather than speakers.[171] However, the rise of AI-generated misinformation indicates that there may be a threshold at which affirmative regulatory intervention becomes necessary to preserve the integrity of that marketplace. These realities counsel in favor of a more nuanced approach to defamation law, particularly one that gives heightened attention to defamatory content created using generative AI.

Less restrictive regulatory approaches that have the potential to survive strict scrutiny include narrowly targeting AI-generated content used in

168. *See, e.g.*, Fair Hous. Council of San Fernando Valley v. Roommates.com, 521 F.3d 1157, 1168 (9th Cir. 2008) (noting that a platform may not be immune from liability where it "contributes materially" to its alleged unlawfulness).

169. *See* Monitor Patriot Co. v. Roy, 401 U.S. 265, 274–75 (1970).

170. *See, e.g.*, Jack Dunn, *The White House Posts AI-Generated Image of Trump as a Buff Jedi: 'Happy May the 4th to You All,'* VARIETY (May 4, 2025, at 12:00 PT), https://variety.com/2025/film/news/ai-generated-image-trump-buff-jedi-star-wars-day-1236386522; Max Matza, *Trump Criticized After Posting AI Image of Himself as Pope*, BBC NEWS (May 4, 2025), https://www.bbc.com/news/articles/cdrg8zkz8d0o; Rachel Hall, *'Trump Gaza' AI Video Intended as Political Satire, Says Creator*, GUARDIAN (Mar. 6, 2025), https://www.theguardian.com/technology/2025/mar/06/trump-gaza-ai-video-intended-as-political-satire-says-creator; Naomi Lachance, *White House Mocks Man Trump Illegally Deported with Obama Campaign Theme*, ROLLING STONE (May 3, 2025), https://www.rollingstone.com/politics/politics-news/trump-white-house-mocks-abrego-garcia-obama-campaign-theme-1235330966.

171. Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

commercial or political advertising, as several states have already begun experimenting with. Rather than displacing these efforts, future regulation should build on and refine such approaches by shifting disclosure obligations away from individual users and toward media platforms, which are better positioned to verify and label certain categories of AI-generated content at scale. Critically, any such regulation must focus on particularly harmful and deceptive conduct, rather than speech content in the abstract. In the context of AI-generated libel against public officials, "harmful" should be narrowly defined as conduct that undermines the government's compelling interest in transparency and the integrity of the electoral process. Policies should target deceptive practices intended to manipulate voters or distort elections; rather than attempting to suppress messages outright, disclosure regimes should instead emphasize source transparency. Even distasteful or misleading content may remain constitutionally protected, but transparency-based warnings can mitigate harm without suppressing speech.

In addition, both courts and legislatures should reconsider the contours of intermediary and amplifier liability to ensure that platforms and identifiable actors who meaningfully contribute to the spread of defamatory AI-generated content bear accountability for the resulting harm. Given the speed and scale of viral dissemination in the digital political arena, time is of the essence. Accordingly, cooperation with social media platforms to facilitate rapid, visible correction mechanisms may be necessary to counteract viral disinformation.[172] Considering existing technological capabilities, AI systems themselves could be deployed to flag and attach warning labels to reposts or reshares of affirmatively identified AI-generated defamatory disinformation. At the same time, any recalibration must preserve *Sullivan*'s core function of protecting robust political debate and preventing self-censorship, recognizing that the doctrine's foundational goals are best served not by overruling the standard, but by adapting its application to contemporary technological realities.

CONCLUSION

Sullivan is a product of its time, crafted to protect a fragile press and to safeguard democratic dissent in a media environment defined by centralized journalism and textual speech. In doing so, the Court heavily relied on a core set of assumptions: (1) that public officials possess meaningful capacity to counter falsehoods, (2) that retractions and corrections are feasible and can effectively repair reputational harm, and (3) that audiences can rationally discern fact from

172. *See, e.g.*, Ethan Porter & Thomas J. Wood, *Political Misinformation and Factual Corrections on the Facebook News Feed: Experimental Evidence*, 84 J. POL. 1812, 1814 (2022) (finding that fact-checks displayed on a simulated Facebook news feed increased user accuracy in identifying misinformation); Yuwei Chuai, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini & Nicholas Pröllochs, *Community-Based Fact Checking Reduces the Spread of Misleading Posts on Social Media* (Sep. 13, 2024) (unpublished manuscript) (discussing X's use of "Community Notes," a fact-checking system aimed at allowing viewers to assess the accuracy of posts).

satire. In an information ecosystem dominated by generative AI, deepfakes, and algorithmic amplification, these assumptions are not merely strained—they are structurally undermined. Yet *Sullivan*'s core normative commitment to uninhibited debate remains vital. The challenge is not whether to preserve *Sullivan*, but how to adapt it to the modern AI-integrated landscape. Preserving *Sullivan*'s constitutional legacy, therefore, requires doctrinal adaptation rather than abandonment. Courts and legislatures must confront how AI-generated falsity destabilizes traditional analyses of intent, satire, and reasonableness, and recalibrate defamation doctrine accordingly. Without such recalibration, defamation law risks protecting neither democratic discourse nor reputational integrity in an era in which falsity is faster, more persuasive, and more difficult to undo than ever before.

***