FREE SPEECH IN SOCIAL MEDIA: BALANCING DEMOCRATIC VALUES AND COMBATING FALSE SPEECH

Artículo

Carlos Chévere Lugo*

Those who won our independence believed that the final end of the State was to make men free to develop their faculties... that freedom to think as you will and to speak as you think are means indispensable to the discovery and spread of political truth.
Justice Louis Brandeis[1]

* Carlos Chévere-Lugo is a Puerto Rican lawyer admitted to the Puerto Rico Bar and the United States Federal Court for the District of Puerto Rico. Chévere-Lugo has an L.LM. in American Legal System (2017) and International Criminal Law (2018) from St. Mary's University School of Law, a J.D. (2016) from Pontifical Catholic University School of Law, and a B.A. in Political Science (2012) from the University of Puerto Rico.

Introduction

Free speech is a cornerstone of democratic principles, allowing individuals to express their opinions, challenge authority, and contribute to collective decision-making. In a democracy, citizens are not merely subjects of governance but active participants whose diverse viewpoints shape public policy and societal norms.[2] The protection of free speech ensures that ideas can be debated and discussed openly, no matter how controversial or unpopular. This exchange of ideas fosters accountability, transparency, and a more informed electorate, all essential for a healthy democracy.[3]

Social media has revolutionized how free speech is exercised, transforming it into a global public square where discourse happens in real-time.[4] Platforms such as Twitter, Facebook, and Instagram allow individuals to voice their opinions to an audience of millions and democratize access to information, breaking down geographical and societal barriers.[5] This shift has given marginalized voices a platform and reshaped traditional power dynamics in communication.[6] Social media's immediacy and reach have made it an indispensable tool for organizing movements, influencing political campaigns, and challenging mainstream narratives.[7]

The transformative role of social media in amplifying free speech also comes with challenges. The vast reach of these platforms can magnify misinformation, hate speech, and extremist views, raising questions about the limits of free speech in a digital age.[8] While democracies value the free exchange of ideas, balancing this freedom with the need for responsible communication is a growing concern.[9] The responsibility of regulating speech on social media often falls into a gray area between governments,

---

1 *Whitney v. California*, 274 U.S. 372, 375 (1927) (Brandeis, J., concurring).

2 *See e.g.,* C. Edwin Baker, *Is Democracy a Sound Basis for a Free Speech Principle?,* 97 Va. L. Rev. 515, 518-19 (2013).

3 *See e.g.,* Helen Norton & Toni M. Massaro, *Free Speech and Democracy: A Primer for Twenty-First Century Reformers,* 54 UC Davis L. Rev. 1631, 1634-35 (2021).

4 Zolbrys, *The Challenges Social Media Poses for Current Free Speech Jurisprudence,* Harvard Civil Rights – Civil Liberties Law Review (Nov. 6, 2018), https://journals.law.harvard.edu/crcl/the-challenges-social-media-poses-for-current-free-speech-jurisprudence/.

5 Greta Filor, *Free Speech in the Age of Social Media,* Brown Undergraduate Law Review (May 30, 2023), https://www.brownulr.org/blogposts/free-speech-in-the-age-of-social-media.

6 *See e.g.,* Jamillah Bowman Williams, et al., *#BlackLivesMatter—Getting from Contemporary Social Movements to Structural Change*, 12 Calif. L. Rev. Online 1, 3 (2021).

7 Platforms like Facebook and Twitter have been instrumental in mobilizing individuals for social and political causes. The Arab Spring is a notable example of how social media facilitated the coordination of protests and the dissemination of information, enabling rapid mobilization across diverse populations. *See* Catherine O'Donnell, *New Study Quantifies Use of Social Media in Arab Spring,* Univ. of Wash. News *(*Sept. 12, 2011), https://www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/; *See e.g.,* Killian Clarke & Korhan Kocak, *Launching Revolution: Social Media and the Egyptian Uprising's First Movers,* 50 Brit. J. Pol. Sci. 1025 (2020).

8 *See Social Media, Freedom of Speech, and the Future of Our Democracy,* Columbia News (last visited February 20, 2025), https://news.columbia.edu/content/social-media-freedom-speech-and-future-our-democracy.

9 *See e.g.,* Social Media, Freedom of Speech, and the Future of Our Democracy 15 (Lee C. Bollinger & Geoffrey R. Stone eds., 2022).

corporations, and users, each grappling with maintaining open discourse while protecting individuals from harm.[10]

Despite the challenges posed by social media, the centrality of free speech to democratic principles remains unchanged, and social media continues to play a pivotal role in shaping the future of public discourse. As digital platforms evolve, the debate surrounding free speech, censorship, and regulation will likely intensify.[11] However, the fundamental idea that free speech is critical for democratic participation, accountability, and societal progress endures. For all its complexities, social media remains one of the most potent tools for exercising and defending this right today.[12]

Free speech on social media upholds democratic ideals by fostering an open exchange of ideas, allowing individuals to voice opinions, share information, and engage in public debate. It empowers marginalized groups, challenges entrenched power structures and promotes civic participation by enabling people to mobilize around causes and push for social or political change.[13] Social media's decentralized nature helps ensure that no single entity controls the narrative, allowing a pluralism of perspectives to flourish. This inclusivity aligns with the democratic principle of ensuring that all citizens can participate in shaping the direction of their society.[14]

However, the same freedom that promotes democratic engagement can also undermine it, primarily through the unchecked spread of misinformation.[15] False or misleading information, often spread rapidly across social media, can distort public understanding of critical issues and erode trust in democratic institutions. Misinformation can polarize societies, fuel conspiracy theories, and manipulate electoral processes, effectively weak-

---

10 Aliza Chasan, et al., *Balance between fighting misinformation and protecting speech on social media gets more complicated*, CBS News (Mar. 24, 2024), https://www.cbsnews.com/news/social-media-misinformation-supreme-court-free-speech-60-minutes/.

11 Valerie C. Brannon, Free Speech and the Regulation of Social Media Content 2019, https://crsreports.congress.gov/product/pdf/R/R45650 (noting the difficulty in balancing First Amendment rights with the necessity of moderating content to prevent harm, highlighting the nuanced legal landscape).

12 Isaac Chotiner, *The Evolving Free-Speech Battle Between Social Media and the Government*, The New Yorker (Jul. 15, 2023), https://www.newyorker.com/news/q-and-a/the-evolving-free-speech-battle-between-social-media-and-the-government.

13 Rajvinder Sunner, *Social Media: How it provided the voiceless with a voice,* NUHA Foundation (Oct. 1, 2019), https://nuhafoundation.org/home/blog/bloggingentries/2019/youth/social-media-how-it-provided-the-voiceless-with-a-voice/ (noting that the #MeToo movement enabled survivors of sexual harassment to share their stories widely, fostering a global conversation on gender-based violence.); John Wihbey, *How does social media use influence political participation and civic engagement? A meta-analysis,* The Journalist's Resource (Oct. 18, 2015), https://journalistsresource.org/politics-and-government/social-media-influence-politics-participation-engagement-meta-analysis/ (noting that social media use is positively associated with various forms of civic engagement, including volunteering, attending political meetings, and voting).

14 Molly Mastantuono, *Why Social Media is a Source of Strength for Black Americans,* Bentley Univ. News (Feb. 24, 2023), https://www.bentley.edu/news/why-social-media-source-strength-black-americans (noting that the Black Lives Matter movement, for example, utilized social media to highlight systemic racism and mobilize global protests).

15 Gabriel R. Sanchez & Keesha Middlemass, *Misinformation is eroding the public's confidence in democracy,* Brookings (Jul. 26, 2022), https://www.brookings.edu/articles/misinformation-is-eroding-the-publics-confidence-in-democracy/.

ening the informed electorate that democracy relies upon.[16] While social media amplifies voices, it also blurs the line between credible and unreliable sources, complicating efforts to maintain a well-informed public, which is crucial for the health of any democracy.[17]

The *Pizzagate* conspiracy theory is a salient case study in the dynamics of misinformation dissemination via social media. Originating in 2016, this false narrative alleged the existence of a child sex-trafficking operation involving prominent political figures, supposedly run out of a pizzeria in Washington, D.C. Initially, the theory gained traction on niche forums such as 4chan and Reddit, leveraging the existing culture of conspiracy theories.[18] Its transition to mainstream platforms like Twitter, Facebook, and Instagram marked a significant escalation in its reach. Utilizing viral hashtags, sensationalized headlines and memes optimized for engagement facilitated a broader audience engagement, with varied motives among users ranging from political defamation to genuine belief in unfounded claims.[19] The amplification of Pizzagate was further propelled by influential right-wing figures and dedicated conspiracy-oriented YouTube channels, all buoyed by platform algorithms favoring content that drives interaction over factual accuracy. This trajectory culminated in a real-world act of violence, exemplified by the armed invasion of a pizzeria based on baseless conspiracy theories, thus demonstrating the tangible and dangerous repercussions of unchecked online disinformation.[20]

This case underscores the critical implications of social media architecture, which prioritizes sensationalism and user engagement, thereby exacerbating the spread of hazardous misinformation. QAnon emerged as a far-reaching conspiracy theory in 2017, rapidly evolving into a digital movement that propagated many baseless claims and misinformation. Initially centered around the idea that a secret cabal of elite figures controlled global politics and that Donald Trump was fighting to dismantle this network, QAnon relied heavily on social media platforms like Twitter, Facebook, and YouTube to amplify its narratives.[21] Anonymous posts from a figure known as *Q* spread coded messages, which followers interpreted and shared widely through hashtags, memes, and viral videos. QA-

---

16 Beata Martin-Rozumilowicz & Rasťo Kužel, *Social Media, Disinformation, and Electoral Integrity: IFES Working Paper*, The International Foundation for Electoral Systems (Aug. 14, 2019), https://www.ifes.org/sites/default/files/migrate/ifes_working_paper_social_media_disinformation_and_electoral_integrity_august_2019_0.pdf.

17 Judy Woodruff & Connor Seitchik, *Social media's role in fueling extremism and misinformation in a divided political climate*, PBS (Sept. 11, 2024), https://www.pbs.org/newshour/show/social-medias-role-in-fueling-extremism-and-misinformation-in-a-divided-political-climate.

18 Mike Wendling, *The saga of 'Pizzagate': The fake story that shows how conspiracy theories spread*, BBC News (Dec. 2, 2016), https://www.bbc.com/news/blogs-trending-38156985.

19 Brian Stelter, *Fake news, real violence: 'Pizzagate' and the consequences of an Internet echo chamber*, CNN Business (Dec. 6, 2016), https://money.cnn.com/2016/12/05/media/fake-news-real-violence-pizzagate/index.html.

20 *'Pizzagate' gunman sentenced to four years*, BBC News (June 22, 2017), https://www.bbc.com/news/world-us-canada-40372407.

21 Mike Wendling, *QAnon: What is it and where did it come from?*, BBC News (Jan. 6, 2021), https://www.bbc.com/news/53498434; Brandy Zadrozny & Ben Collins, Jance C. Timm, Trump Pushed QAnon and 4chan-Created Conspiracy Theories in Georgia Call, NBC News (Jan. 4, 2021), https://www.nbcnews.com/tech/internet/trump-pushed-qanon-4chan-created-conspiracy-theories-georgia-call-n1252769; Brett Fujioka, *Toxic Internet Culture from East to West*, Noema (Oct. 22, 2020), https://www.noemamag.com/toxic-internet-culture-from-east-to-west/.

non's flexible, open-ended messaging adapted to various conspiratorial themes, including COVID-19 vaccine misinformation, election fraud claims, and other false narratives.[22] The movement leveraged algorithm-driven amplification and community-building tools to quickly and efficiently recruit followers. This strategy not only helped QAnon gain a global following but also contributed to real-world harm, with some followers engaging in violent acts, disrupting public health efforts, and participating in the January 6 Capitol attack.[23] By exploiting the design of social media platforms that prioritize engagement, QAnon demonstrated how digital networks can be weaponized to spread misinformation at scale.

In this article, we explore the evolving role of free speech on social media, its challenges, and the legal implications of balancing democratic values with the rise of misinformation. Part I examines free speech's historical and philosophical roots, and the importance of protecting even false speech within democratic societies. Part II discusses the implications of online misinformation on democratic stability, highlighting key incidents and the damaging effects of unchecked false information. Part III delves into the practices and legal challenges of content moderation on social media, contrasting global regulatory approaches like the U.S.' Communications Decency Act and the European Union's Digital Services Act. Part IV proposes potential policy frameworks to responsibly manage misinformation without infringing on fundamental free speech rights.

I. False Speech and the Role of Free Speech in a Democracy

A. Historical Foundations of Free Speech

The concept of free speech has a rich history, dating back to ancient Greece, where it was viewed as a fundamental aspect of democratic society. In classical Athens, the notion of *parrhesia*, or the right to speak freely, was crucial to the functioning of democracy.[24] Citizens were urged to voice their opinions in the Assembly, the political center where laws were deliberated and decisions made.[25] While male citizens primarily enjoyed this liberty, it represented a significant philosophical and cultural shift towards appreciating individual expression in governance.[26] Philosophers such as Socrates, known for challeng-

---

22 *See Id.*; Ian Haimowitz, *No One Is Immune: The Spread of Q-Anon Through Social Media and the Pandemic*, Center for Strategic & International Studies (Dec. 17, 2020), https://www.csis.org/blogs/strategic-technologies-blog/no-one-immune-spread-q-anon-through-social-media-and-pandemic.

23 *See* Wendling, *supra* note 21; Haimowitz, *supra* note 22; *Capitol riot: 'QAnon Shaman' Jacob Chansley sentenced to 41 months in prison*, BBC News (Nov. 17, 2021), https://www.bbc.com/news/world-us-canada-59253090.

24 David Konstan, *The Two Faces of Parrhêsia: Free Speech and Self-Expression in Ancient Greece*, 46 Antichthon 1, 10 (2012); *See* Arlene W. Saxonhouse, *Parrhêsia: The Unbridled Tongue in Ancient Democratic Athens*, *in* Populism, Demagoguery, and Rhetoric in Historical Perspective 33, 35-50 (Giuseppe Ballacci & Rob Goodman eds., Oxford University Press 2023).

25 *See* Jacob Mchangama, Free Speech: A History from Socrates to Social Media 13 (Basic Books, 2022). ("The Athenians had two distinct but overlapping concepts of free speech. *Isēgoría* refers to equality of the public, and civic speech, while *parrhēsía* can be translated as 'frank' or 'uninhibited' speech. *Isēgoría* was exercised in the Athenian Assembly-the *ekklēsia,* where each session opened with the question, *'w*ho wishes to speak?' *Parrhēsía* allowed citizens to be bold and honest in expressing their opinions even when outside the assembly and extended to many spheres of Athenian life including philosophy and theater.").

26 *Id.* ("The nineteenth-century English historian and radical member of Parliament George Grote, who did much to rehabilitate the Athenian democracy as a model for liberal reform movements, emphasized 'the liberty

ing traditional beliefs, exemplified the significance and dangers of this freedom, as he was ultimately condemned to death for allegedly corrupting the youth and showing disrespect to the gods.[27]

Ancient Rome inherited many democratic principles from Greece but adapted them to fit its unique context. While the Roman Republic allowed some degree of free speech in its Senate, it was more restricted than in Athens.[28] The concept of *libertas*, or liberty, was celebrated, but speaking out against those in power could be perilous.[29] The transition from Republic to Empire further eroded free speech, as emperors, like Augustus, and rulers implemented laws against *maiestas* (treason), targeting dissenters and critics.[30] For instance, the case of Cassius Severo, who was exiled for his writings, demonstrates the precarious nature of expression under autocratic rule.[31]

During the early medieval period, the concept of free speech faced significant challenges, particularly with the rise of the Inquisition. As Christianity became the dominant force in Europe, the Church sought to control religious practices and the flow of ideas. The early Inquisition, which began in the 12th century, was primarily aimed at rooting out heresy.[32] Speaking against the doctrines of the Church or advocating unorthodox beliefs could result in severe punishment, including excommunication,

---

of thought and action at Athens, not merely excessive restraint of law, but also from a practical intolerance between man and man, and tyranny of the majority over individual dissenters in taste and pursuit.'").

27 *See* Arlene W. Saxonhouse, Free Speech and Democracy in Ancient Athens 7 (Cambridge University Press 2006) (Socrates' trial in 399 BCE highlights the complexities surrounding free speech in Athens. Accused of impiety and corrupting the youth, his conviction and execution underscore the limitations and potential dangers of exercising free speech in a democratic society.).

28 *See* Keith Werhan, *The Classical Athenian Ancestry of American Freedom of Speech*, 2008 Sup. Ct. Rev. 293, 300 (2008) ("As practiced in the classical Athenian democracy, the principle of isēgoria offered every full citizen in good standing an equal opportunity to make proposals and to speak before the Assembly (ekklēsia)."); Mchangama, *supra* note 25, at 24. ("Yet there were no Roman equivalents of the Greek terms isēgoría and *parrhēsía*. Roman free speech was first and foremost exercised in the Senate, by magistrates before assemblies, and by orators before the courts, where, as in Athens, political speech would often be interwoven with legal arguments. For men like Cicero and Caesar, oratory was an essential way to further their political careers. Had Caesar not been a brilliant orator, he may not have become a brilliant general—or dictator.").

29 Mchangama, *supra* note 25 at 24. ("The Romans contrasted *libertas* with *licentia*, or 'licentiousness.' *Licentia* was essentially an abuse of freedom that was either illegal or very much frowned upon. The ancient Roman historian Tacitus wrote that, in contrast to Rome, with the Greeks 'not the freedom only, but even the licentiousness of speech, is unpunished.").

30 *Id*. at 27. *See, e.g.,* Karl Galinsky, *Freedom of Speech in the Reign of Augustus: How Much of an Issue?*, 53 Arethusa 247 (2020); Thomas E. Strunk, On the Fall of the Roman Republic 77-80 (Anthem Press, 2022) (Kindle e-book).

31 Mchangama, *supra* note 25, at 27. ("[T]he outspoken orator Cassius Severus and his insolent writings against a number of prominent Romans provoked Augustus to widen the scope of the law to include words and writings. Severus was convicted and banished to Crete while his entire writings were burned. But this was merely the beginning of a purge. What historian Frederick H. Cramer has called 'literary treason'—purely verbal or written attacks on the government—was now a punishable crime. Charged with this new crime, teachers were taken to court and more writers saw their entire life's work go up in flames. But Augustus did not stop there. The historian Suetonius implies that the Senate made it illegal to own, circulate, or even read the writings of a condemned author.").

32 *Id*. at 50. ("Heresy thus became one of the defining issues of the thirteenth and fourteenth centuries just as pagan teachings started enjoying a prominent role at the newly emerged universities where reason, inquiry, and science thrived. The quest to eradicate heresy reshaped Western Europe into what British historian R. I. Moore has called a 'persecuting society.'").

imprisonment, or execution.[33] This period marked a stark departure from the relatively open discourse of classical antiquity, as the Church's authority overrode individual expression.

Despite these restrictions, the medieval period also saw the emergence of spaces where free speech could flourish in limited forms. Universities began to develop in the 11th and 12th centuries and became centers of learning and debate.[34] Although theological orthodoxy was paramount, scholars engaged in philosophical and scientific discussions that laid the groundwork for the Renaissance.[35]

In late medieval times, developing nation-states resisted allowing freedom of speech in public forums, regardless of the content. This resistance resulted from the authoritarian government structure of that time. Religious authority justified political authority, and truth was determined by divine revelation. Disagreement with this authority was not just an error but also considered a severe offense. Opposing opinions and dissent became less tolerated when governments started to address political rather than religious needs. Justifying taxes and military conscription required the support of the population. Lord Holt, in *Queen v. Tutchin*, expressed:

> If men should not be called to account for possessing the people with an ill opinion of the Government, no Government can subsist; for it is very necessary for every Government, that the people should have a good opinion of it. And nothing can be worse to any Government, than to endeavor to produce animosities as to the management of it.[36]

In England, the situation was further complicated by the conflict between the Roman Catholic and English Catholic churches and the protracted power struggles between the king and parliament. In the three centuries preceding the Declaration of Independence, winning the hearts and minds of the English people entailed suppressing opposing ideas.

---

33 *Id*. at 52. ("In 1184, Pope Lucius III perpetually anathematized 'all who presume to think, or to teach … otherwise than as the Holy Roman Church teaches and observes.' He ordered every bishop to annually comb his parish for heretics and punish them accordingly. In 1199, Pope Innocent III declared heresy to be a crime of treason against God himself and he treated it as such. Not only were heretics to have their belongings confiscated, even their children were damned to a life of poverty.").

34 James Hankins, *Intellectual Freedom in Medieval Universities*, First Things (Feb. 4, 2022), https://www.firstthings.com/web-exclusives/2022/02/intellectual-freedom-in-medieval-universities (The development of medieval universities was influenced by the need to regulate intellectual liberty and address heterodox thought. The earliest universities, such as Bologna, Cambridge, Paris, and Oxford, were established between 1190 and 1230, partly in response to European heretical movements. These institutions aimed to transmit learning and maintain orthodoxy rather than encourage free thought.).

35 *See* Paul Meany, *The Medieval Case for Free Speech,* We Speak Freely (Aug. 25, 2018), https://www.wespeakfreely.org/2018/08/25/medieval-case-free-speech (Figures like John of Salisbury advocated for a form of free speech, emphasizing the role of debate in virtuous and just governance. His views reflect an early recognition of the importance of open discourse in shaping society—a sentiment echoed in the academic culture of medieval universities, where controlled yet meaningful debate was possible.).

36 Patti Goldman, *Combatting the Opposition: English and United States Restrictions on the Public Right of Access to Governmental Information,* 8 Hastings Int'l & Comp. L. Rev. 249, 256 (1985) (*citing* Queen v. Tutchin, 14 State Trials 1095, 1133-34 (Q. B. 1704) (Holt, C.J.).).

The primary methods of achieving this suppression were through seditious defamation laws and press regulation via licensing.[37]

The initial migration to the New World is often attributed to the repressive religious policies that were prevalent in Europe at that time. Many who escaped religious persecution quickly established settlements that discriminated against non-believers. The decline in censorship as a tool in England undoubtedly contributed to its absence in the pre-revolutionary era. However, the doctrine of seditious libel was in full force on both sides of the Atlantic Ocean.[38] Although the last seditious libel trial in the colonies, involving the New York printing press owned by John Peter Zenger, occurred in 1735, the threat of censorship and prosecution persisted.[39]

The trial of John Peter Zenger marked a watershed moment in American legal history as the first famous free speech and press case. Zenger had published articles criticizing New York Governor William Crosby's public policies, leading to his prosecution.[40] However, his defense, led by prominent lawyer Andrew Hamilton, argued that truth was the best defense against a criminal prosecution for seditious defamation. Hamilton's persuasive argument ultimately led the jury to deliver a verdict of not guilty in just 10 minutes. Despite the trial's immediate impact, the concept of freedom of press and speech continued to evolve until the adoption of the First Amendment in 1792.

During the colonial era, politicians outwardly championed the right to criticize the government freely. However, behind closed doors, they manipulated and censored speech to better navigate the political landscape of the time. While issues surrounding freedom of expression were primarily intertwined with the general debate on the revolution, the guiding principles of common law during the drafting of the Constitution were already eloquently articulated by the English jurist William Blackstone. Although censorship had significantly waned in England during that period, seditious libel remained widespread. As Blackstone put it:

> The liberty of the press is indeed essential to the nature of a free state; but this consists in laying no previous restraint upon publications, and not in freedom from censure for criminal matter when published. Every freeman has an undoubted right to lay what sentiments he pleases before the public; to forbid this, is to destroy the freedom of the press; but if he publishes what is improper, mischievous, or illegal, he must take the consequences of his own temerity . . . Thus, the will of individuals is still left free; the abuse only of that free-will is the object of legal punishment. Neither is any restraint hereby laid upon freedom of

---

37 GORDON S. WOOD, EMPIRE OF LIBERTY: A HISTORY OF THE EARLY REPUBLIC, 1789–1815 258-59 (Oxford Univ. Press 2009); Christopher D. Jenkins, *The Sedition Act of 1798 and the Incorporation of Seditious Libel into First Amendment Jurisprudence* 11 (M.A. thesis, Marshall Univ. 2013), https://mds.marshall.edu/cgi/viewcontent.cgi?article=2688&context=etd.

38 *See e.g.,* Harold L. Nelson, *Seditious Libel in Colonial America*, 3 AM. J. LEGAL HIST. 160, 166 (1959).

39 *See Id.* at 164, 170; Eleanor Stratton, *John Peter Zenger: Press Freedom Debate*, U.S. CONSTITUTION.NET (Aug. 15, 2024), https://www.usconstitution.net/john-peter-zenger-press-freedom-debate/.

40 *See* MICHAEL S. ARIENS, AMERICAN CONSTITUTIONAL LAW AND HISTORY 655 (Carolina Academic Press 2nd ed. 2016).

> thought or enquiry: liberty of private sentiment is still left; the disseminating, or making public, of bad sentiments, destructive of the ends of society, is the crime which society corrects.[41]

Blackstone aimed to outline the general common law doctrine at his time regarding freedom of speech and press. Professor Wendell Bird's study on freedom of expression during Blackstone's era suggests a more extensive and robust interpretation. After thoroughly examining English texts from Blackstone's time, the professor concludes:

> [T]hat the dominant understanding of liberties of press and speech in preserved publications, from the 1770s to the early 1790s, was an expansive meaning of those liberties and a belief that they conflicted with criminalizing seditious libel. Only a declining minority, led by Crown judges, viewed those liberties in the narrow way Blackstone purported to summarize, and that [Leonard] Levy ascribed to the supporters of Fox's Libel Act and to the framers of the First Amendment.[42]

Upon ratifying the First Amendment in 1789, the crime of seditious defamation was not eradicated. The fiercely contested 1796 presidential election between John Adams of the Federalist Party, and Thomas Jefferson of the Republican Party—now the Democrat—resulted in Adams winning, yet the hatred endured. In 1798, the Federalist-controlled Congress enacted the Alien and Sedition Act, which included a provision prohibiting any false, scandalous, and malicious writing.[43] Utterances seeking to defame, disparage, or discredit were deemed criminal. While malicious intent was a crucial aspect of the offense, the truth did not serve as a defense. The Adams administration brought charges against several of Jefferson's supporters for violating the Sedition Act.[44]

Following the 1800 election, President Thomas Jefferson issued pardons to all individuals convicted under the Sedition Acts and decided against pursuing additional prosecutions under that law. Subsequently, a newly elected, Republican-controlled Congress passed legislation to compensate fines imposed on those found guilty under the Sedition Act.[45] From the expiration of the Sedition Acts in 1800 until 1917, the Free Speech Clause of the First Amendment was seldom invoked.[46] This is because the clause applied only to *Congress*, and the incorporation of the Bill of Rights into the Due Process Clause of the

---

41 *Freedom of Expression - Speech and Press*, JUSTIA U.S. Law (last visited Feb. 21, 2025), https://law.justia.com/constitution/us/amendment-01/04-freedom-of-expression-speech-and-press.html (*citing* 4 William Blackstone, Commentaries on the Laws of England 151-52 (T. Cooley ed., 2d ed. rev. 1872)).

42 Wendell Bird, *Liberties of Press and Speech: 'Evidence Does Not Exist To Contradict the…Blackstonian Sense' in Late 18th Century England?*, 36 Oxford J. Legal Stud. 1-2 (2016), https://www.jstor.org/stable/26363954.

43 *See, e.g.*, Charles Slack, Liberty's First Crisis: Adams, Jefferson, and the Misfits Who Saved Free Speech (Atlantic Monthly Press, 2015).

44 *See* Michael P. Downey, *The Jeffersonian Myth in the Supreme Court Sedition Jurisprudence*, 76 Wash. U. L.Q. 683, 707 n.59 (1998) ("Twenty-five people were prosecuted and ten convicted under the Sedition Act. . . Typical cases prosecuted under the Sedition Acts were the case of editor Thomas Cooper, who was sentenced to six months imprisonment and a $400 fine, and the case of publisher James Callender, who received a nine month prison term and a $200 fine.").

45 *Id.* at 693-94.

46 *See, e.g.*, David M. Rabban, Free Speech in Its Forgotten Years, (Cambridge University Press 1997).

Fourteenth Amendment had not yet taken place.[47] Another significant reason for the lack of use of the free speech clause was that Congress passed relatively few speech-related laws.

B. Why Free Speech?

Political and legal philosophers have presented several reasons in support of free speech. First, speech is essential for safeguarding individual liberty from governmental tyranny, as exemplified in the *Zenger* case. Without freedom of speech and free press, tyrants would go unchecked. In this context, freedom of speech and the press check the government's exercise of power, and these rights are integral to self-government in a democratic society.[48] As Justice O'Connor stated, political speech regulations strike at the core of the First Amendment.[49]

Second, discourse, as a tool for helping people seek truth on various issues, plays a significant role in promoting free speech. Human fallibility and the difficulty of discovering the truth make the right to free speech a necessity rather than a luxury. As Judge Oliver Wendell Holmes, Jr.'s dissenting opinion states, encouraging freedom of expression increases the likelihood of uncovering the truth, as speech competes in a marketplace of ideas.[50]

Thirdly, freedom of speech exists to enable individuals to pursue their fulfillment. Discourse is not just a tool for political self-government or truth discovery but an end in itself. Speech and expression would allow people to achieve their self-set goals. This self-actualization rationale is a contemporary addition to philosophical justifications for free speech.

C. Free Speech and the Digital Age

Social media has played a transformative role in reshaping public discourse, particularly for voices historically marginalized by mainstream media.[51] Platforms such as Twitter, Instagram, and TikTok have empowered individuals to bypass traditional gatekeepers, like editors or broadcasters, who previously controlled the topics that received attention. This shift has allowed marginalized groups—whether based on race, gender, class, or sexual orientation—to share their perspectives with a global audience.[52] Unlike mainstream media, where commercial interests often influence coverage, social media provides a more direct and unfiltered channel for expression.[53] The viral nature of social media ensures

---

47 *See* Michael S. Ariens, *supra* note 40.

48 *Id.*

49 Boos v. Barry, 485 U.S. 312, 318 (1988).

50 Abrams v. United States, 250 U.S. 616, 630 (1919).

51 *See* Jeremy Robinson, *The Modern Public Forum: Theoretical Liability to Aid in Protecting Free Speech on Social Media Platforms*, 2 No. 3 Md. B.J. 102 (2021) (exploring the evolving role of social media platforms as contemporary public forums and the legal implications of this transformation, particularly in the context of free speech protections and liability).

52 *See Section 230 as First Amendment Rule*, 131 Harv. L. Rev. 2027 (2018) (discussing how social media platforms have become essential venues for free expression, particularly for marginalized communities).

53 *See* Olivier Sylvain, *Platform Realism, Informational Inequality, and Section 230 Reform,* 131 Yale L.J. Forum 475, 478-79 (2021) (highlighting how these platforms allow marginalized groups to disseminate information without traditional intermediaries, thereby challenging existing power structures in media).

that these stories can spread rapidly, reaching audiences far beyond what traditional media could achieve.

Furthermore, social media provides tools for real-time communication, making it easier for marginalized voices to respond to current events and shape narratives as they unfold. For example, during political unrest or social injustice, individuals on site can report incidents as they happen, offering perspectives that might otherwise be ignored or misrepresented by mainstream outlets.[54] This ability to report in real-time also gives rise to citizen journalism, where everyday people play a crucial role in shaping public opinion.[55] By creating alternative narratives, marginalized groups can challenge dominant media portrayals and offer a more nuanced view of issues affecting them.

Another critical aspect of social media's role in amplifying marginalized voices is the creation of spaces for solidarity and support.[56] Many marginalized individuals find communities online where they can express themselves freely and connect with others who share similar experiences. These online spaces often provide validation and support in a world where mainstream narratives can be exclusionary or discriminatory.[57] Additionally, while not without flaws, social media algorithms can sometimes help individuals discover these communities, further enhancing their ability to participate in public discourse.

D. Philosophical Foundations for the Protection of False Speech

The question of whether to protect false speech has been a topic of philosophical discussion for many years. In his work *On Liberty*, John Stuart Mill advocates for the protection of false opinions, arguing that they play a crucial role in the pursuit of truth.[58] Mill contends that truth is not self-evident and must be discovered through debate and discussion. He believes that false ideas can prompt individuals to critically examine their beliefs, ultimately leading to a clearer understanding of the truth.[59] According to Mill, suppressing misguided ideas would hinder this discovery process.[60] For him, the ongoing defense and articulation of truth in the face of opposing views reaffirms its validity and enhances its strength.[61] In this way, free speech can furnish

---

54 *See S*aima Khan, *Rise of Alternative Media and Citizen Journalism*, THE GEOSTRATA (Oct. 13, 2023), https://www.thegeostrata.com/post/rise-of-alternative-media-and-citizen-journalism.

55 *See, e.g.,* Nadine Jurrat, *Mapping Digital Media: Citizen Journalism and the Internet*, OPEN SOCIETY FOUNDATION (Jul. 2011), https://www.opensocietyfoundations.org/publications/mapping-digital-media-citizen-journalism-and-internet.

56 *See e.g.*, Michele Estrin Gilman, *Beyond Window Dressing: Public Participation for Marginalized Communities in the Datafied Society*, 91 FORDHAM L. REV. 503, 508, 554-55 (2022).

57 *See, e.g.,* Florian Arendt, *Media stereotypes, prejudice, and preference-based reinforcement: toward the dynamic of self-reinforcing effects by integrating audience selectivity,* 73 J. COMM. 463, 472 (2023).

58 Daniela C. Manzi, *Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News*, 87 FORDHAM L. REV. 2623, 2626 (2019) (*quoting* John Stuart Mill, *On Liberty*, *reprinted in* ON LIBERTY, UTILITARIANISM AND OTHER ESSSAYS 5, 15, 18–54 (Mark Philip & Frederick Rosen eds., 2015)).

59 *Id.*

60 *Id.*

61 *Id. (quoting* John Stuart Mill, *supra* note 58, at 35 ("[I]f [an idea] is not fully, frequently, and fearlessly discussed, it will be held as a dead dogma, not a living truth.")).

a "clearer perception and livelier impression of truth, produced by its collision with error."[62]

Conversely, Immanuel Kant addresses the ethical implications of intentional falsehoods in his essay *On a Supposed Right to Lie Because of Philanthropic Concerns*.[63] Kant asserts that all forms of lying are morally reprehensible because they violate the dignity of others by obstructing their ability to act freely and rationally.[64] When individuals lie, they deprive others of the right to receive truthful information, thereby manipulating their capacity to make well-informed decisions.[65] Moreover, Kant argues that lying erodes trust in communication as a whole, as it diminishes the speaker's credibility and fosters skepticism among individuals.[66] This, in turn, invokes a strong sense of moral duty to avoid deceit and uphold the integrity of communication, thereby contributing to social harmony and mutual respect.

Together, the perspectives of Mill and Kant reflect a comprehensive view of the importance of free speech and truthful communication. Mill emphasizes the value of open discourse and the potential of false opinions to contribute to discovering truth. At the same time, Kant underscores the ethical duty to avoid deception and uphold the integrity of communication. Despite their differing views on false speech, both philosophers ultimately support the right to express and receive information in a way that respects the autonomy and rationality of individuals. These philosophical frameworks highlight the nuanced relationship between free expression and ethical responsibility. While Mill advocates protecting all speech to foster truth, Kant insists on moral obligation to prevent harm through deceit. Balancing these perspectives, each with its own merits and implications, can help inform contemporary discussions on the complex and multifaceted issue of the limits and responsibilities of free speech in society.

U.S. Supreme Court Judge Oliver Wendell, Jr. enshrined the *marketplace of ideas* concept in American free speech legal history. He wrote: "[T]he ultimate good desired is better reached by free trade in ideas—that the best of truth is the power of the thought to get itself accepted in the competition of the market . . . ."[67] According to this perspective, the open exchange of ideas supports a democratic system by enabling citizens to identify the policies that most effectively benefit society. Drawing on the principles of both Mill and Kant, the concept of the marketplace of ideas suggests that individuals can only broaden their understanding when they can express their viewpoints and engage with opposing perspectives.[68] This framework assumes that truth can be discerned through such interactions and that those involved in these discussions are genuinely committed to seeking the truth.[69]

---

62 *Id.*

63 *Id.* (*quoting* Immanuel Kant, *On a Supposed Right to Lie Because of Philanthropic Concerns*, *reprinted in* Ethical Philosophy 162 (James W. Ellington trans., 2d ed. 1994)).

64 *Id.*

65 *Id.*

66 *Id.*

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

68 *See* Manzi, *supra* note 58 (*quoting* Robert C. Post, Democracy, Expertise, Academic Freedom: A First Amendment Jurisprudence for the Modern State 6 (2012)).

69 Manzi, *supra* note 58, at 2627 n.16 (*citing* Mill, *On Liberty*, *reprinted in* On Liberty, Utilitarianism and Other Esssays 5, 21 (Mark Philip & Frederick Rosen eds., 2015) ("It is the duty of governments, and of individuals,

E. Constitutional Analysis of False Speech

i. Introduction

The Free Speech Clause of the First Amendment prohibits the government from "abridging the freedom of speech," yet it does not clearly define what this freedom encompasses.[70] Recently, federal and state legislators have shown interest in regulating online misinformation and disinformation, which could raise concerns under the Free Speech Clause of the First Amendment.[71] The Supreme Court has traditionally interpreted this Clause to significantly restrict governmental regulation of *protected* speech, which includes certain forms of expressive conduct.[72] However, the government is afforded more flexibility in regulating specific categories of speech that the Court has determined to be *unprotected*.

The Court has consistently affirmed that the First Amendment protects individual and collective speech across a broad spectrum of political, social, economic, educational, religious, and cultural activities. Generally, speech is protected under the First Amendment unless it falls into one of the limited categories of unprotected speech. These unprotected categories, as identified by the Court, include obscenity,[73] defamation,[74] fraud,[75] incitement,[76] fighting words,[77] true threats,[78] speech integral to criminal conduct,[79] and child pornography.[80] Over time, the scope of these categories has been refined, with the Court often narrowing their application. Moreover, the Court has resisted expanding this list to include new categories, such as violent entertainment or depictions of animal cruelty.[81]

The Free Speech Clause of the First Amendment in the United States provides broad protections, including for hate speech; making it one of the few jurisdictions in the world where such speech is legally allowed.[82] In contrast to many other countries that have laws

---

to form the truest opinions they can; to form them carefully, and never impose them upon others unless they are quite sure of being right.")) (Like Kant's belief that people have a moral duty to tell the truth, Mill believed that people have a moral duty to assert only the opinions they sincerely believe.).

70 U.S. Const. amend. I. ("Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.").

71 *See e.g.,* Educating Against Misinformation and Disinformation Act, H.R. 6971, 117th Cong. § 101 (2022), https://www.congress.gov/bill/117th-congress/house-bill/6971; Countering Foreign Propaganda and Disinformation Act, S. 3274, 114th Cong. § 2 (2016), https://www.congress.gov/bill/114th-congress/senate-bill/3274.

72 *See* Texas v. Johnson, 491 U.S. 397 (1989) (holding that burning the American flag as a form of political protest is protected expressive conduct under the First Amendment.); Spence v. Washington, 418 U.S. 405 (1974) (holding that displaying an American flag upside down with a peace symbol was protected expressive conduct.).

73 Miller v. California, 413 U.S. 15 (1973).

74 *See e.g.*, Samantha Barbas, Actual Malice: Civil Rights and Freedom Of The Press In New York Times v. Sullivan (University of California Press 2023).

75 Donaldson v. Read Magazine, Inc., 333 U.S. 178 (1948); Illinois ex rel. Madigan v. Telemarketing Associates, 538 U.S. 600 (2003).

76 Brandenburg v. Ohio, 395 U.S. 444 (1969).

77 Chaplinsky v. New Hampshire, 315 U.S. 568 (1942).

78 Counterman v. Colorado, 600 U.S. 66 (2023).

79 Giboney v. Empire Storage & Ice Co., 336 U.S. 490 (1949).

80 Ashcroft v. Free Speech Coalition, 535 U.S. 234 (2002).

81 United States v. Stevens, 559 U.S. 460 (2010).

82 There has been a lot of published commentary on hate speech and the First Amendment. *See e.g.,* Ronald D. Rotunda, *The Right to Shout Fire in a Crowded Theatre: Hateful Speech and the First Amendment*, 22 Chap.

prohibiting hate speech, the U.S. has consistently maintained that even offensive or harmful speech is protected unless it directly incites violence.[83] This divergence is evident in the United Nations International Covenant on Civil and Political Rights (ICCPR), which prohibits hate speech.[84] However, the U.S. Senate did not ratify provisions of the treaty related to hate speech, reflecting the country's strong commitment to free expression, even when the speech in question is widely regarded as harmful or divisive.[85]

ii. First Amendment Protection of False Speech

Regarding false speech, the Supreme Court has held that, in general, the First Amendment protects false statements unless they belong to a specific unprotected category. While the government can regulate certain subtypes of false speech without violating the First Amendment, this power is limited. The Court has acknowledged that false statements may contribute little to the marketplace of ideas. Nevertheless, there is a concern that regulating false speech could also suppress more valuable forms of expression, leading to self-censorship. Thus, the First Amendment provides *breathing space* for false statements and exaggeration, considered a natural part of free debate, as articulated in *New York Times Co. v. Sullivan*.[86] The Court has suggested that while false ideas may not be regulated, even false factual statements are afforded some constitutional protection.

Generally, laws targeting speech based on its content are subject to strict scrutiny.[87] Under this rigorous standard, a content-based regulation is presumed unconstitutional unless the government can demonstrate that the law is the least restrictive means of

---

L. Rev. 319 (2019); Jeff Kosseff, Liar in a Crowed Theater: Freedom of Speech in a World of Misinformation (John Hopkins University Press 2023); Alexander Brown & Adriana Sinclair, Hate Speech Frontiers: Exploring the Limits of the Ordinary and Legal Concepts (Cambridge Univ. Press 2023) (Kindle e-book); Logan Kline, *You Don't Say: American First Amendment Protection of Hate Speech*, U. Cin. L. Rev. (2021); Audrey Fino, *Defining Hate Speech: A Seemingly Elusive Task*, 18 J. Int'l Crim. Just. 31 (2020); Sophia Staniunas, *Hate Speech: Its Protection Under the First Amendment and Resisting It with Counterspeech*, Fordham Undergraduate Law Review Blog (last visited Feb. 21, 2025), https://undergradlawreview.blog.fordham.edu/first-amendment/hate-speech-its-protection-under-the-first-amendment-and-resisting-it-with-counterspeech/; Alexander Brown & Adriana Sinclair, The Politics of Hate Speech Laws (Routledge 2019); James Weinstein, *Hate Speech Bans, Democracy, and Political Legitimacy*, 32 Const. Comment. 527 (2017); Alex Brown, Hate Speech Law: A Philosophical Examination (2015); Jeremy Waldron, The Harm in Hate Speech (Harvard Univ. Press 2012).

83 Brandenburg v. Ohio, 395 U.S. 444 (1969) (holding that speech advocating illegal conduct is protected unless it is "directed to inciting or producing imminent lawless action and is likely to incite or produce such action.").

84 *See* G.A. Res. 2200 (XXI), International Covenant on Civil and Political Rights ("Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.").

85 *See* Senate Comm. on Foreign Relations, Report on the International Covenant on Civil and Political Rights, S. Exec. Rep. No. 23, 1 (102nd Sess. 1992) ("That Article 20 does not authorize or require legislation or other action by the United States that would restrict the right of free speech and association protected by the Constitution and laws of the United States.").

86 New York Times Co. v. Sullivan, 376 U.S. 254 (1964).

87 *See* Reed v. Town of Gilbert, 576 U.S. 155, 163 (2015) ("Government regulation of speech is content based if a law applies to particular speech because of the topic discussed or the idea or message expressed.").

achieving a compelling governmental interest.[88] This principle has sometimes been applied to statutes regulating false speech, with courts determining that restrictions on lies about specific topics also require strict scrutiny.

The Supreme Court has been divided over the appropriate level of scrutiny for regulating false speech. In *United States v. Alvarez*, the Court struck down the Stolen Valor Act, a federal statute that criminalized false claims about receiving military honors.[89] The plurality opinion of four Justices emphasized that falsity alone is not enough to exclude speech from First Amendment protection.[90] Consequently, the plurality subjected the Stolen Valor Act to strict scrutiny, viewing it as a content-based regulation. The Court found the law overly broad because it penalized false statements without considering their context or intent. There was no *direct causal link* proving that the law's extensive reach was essential to the government's interest in safeguarding the integrity of military awards. In a concurring opinion, two justices found the law unconstitutional but applied a lower, *intermediate* level of scrutiny.[91] This concurrence argued that regulations might meet intermediate scrutiny if they target a "subset of lies where specific harm is more likely to occur."[92]

iii. Constitutional Prohibitions on False Speech

There are instances where restrictions are deemed permissible. For example, the government can impose limits on false statements within specific regulated areas where the Court has permitted content-based restrictions beyond what is constitutionally allowable for fully protected speech. The Court has consistently asserted that false or misleading commercial speech does not fall under the protections of the First Amendment, enabling the government to restrict false statements of fact or mandate additional disclosures in commercial advertising and product labeling to prevent consumer confusion or deception.[93] Fraud is categorized as unprotected speech, encompassing false commercial advertising and broader false representations of facts that deceive or are intended to deceive others, resulting in legal injury.[94]

Agencies like the Federal Trade Commission (hereinafter, "FTC") and the Federal Communications Commission (hereinafter, "FCC") can regulate deceptive commercial speech without violating the First Amendment.[95] For instance, during the COVID-19 pandemic, the FTC issued cease-and-desist letters to companies that falsely claimed that their

---

88 *Id.* ("Content-based laws—those that target speech based on its communicative content—are presumptively unconstitutional and may be justified only if the government proves that they are narrowly tailored to serve compelling state interests.").

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

90 *Id.* at 719.

91 *Id*. at 731 (Breyer, J., concurring) (strict scrutiny leads to "near-automatic condemnation").

92 *Id*. at 736.

93 *See* Virginia. State Bd. of Pharmacy v. Virginia Citizens Consumer Council, Inc., 425 U.S. 748 (1976); Central Hudson Gas & Elec. Corp. v. Pub. Service Commission of New York, 447 U.S. 557 (1980).

94 *Alvarez*, 567 U.S. at 717.

95 *See* Fed. Trade Comm'n, Federal Trade Commission Advertising Enforcement (2008), www.ftc.gov/sites/default/files/attachments/training-materials/enforcement.pdf.

products could treat or prevent COVID-19.[96] However the Court has emphasized that the government cannot bypass First Amendment protections by merely labeling an action as *fraud*.[97] Most recently, the Court has affirmed that many criminal false statement laws, which require proof of intent and are narrowly tailored to avoid harm or significant risk of harm to government entities or processes, are consistent with the free speech guarantee.[98]

Nevertheless, there are limits to the restriction of false speech. In *New York Times v. Sullivan*, the Supreme Court ruled that the Constitution imposes restrictions on common law defamation rules. The Court recognized "[t]hat erroneous statement is inevitable in free debate."[99] To safeguard truthful speech on public issues, the Court extended protections to certain negligent false statements of fact. Under this doctrine, public officials or public figures can only seek damages for reputational harm by proving the statement's falsity and demonstrating with *clear and convincing* evidence that the speaker acted with *actual malice*, indicating they knew the statement was false or acted with reckless disregard for the truth.[100] A similar heightened standard applies to private individuals who become *limited public figures* by engaging in specific public controversies.[101]

The *actual malice* standard also extends to other torts that may result in damages for false speech in public discourse. One such tort is *false light invasion of privacy*, which arises when an individual suffers emotional distress due to highly offensive misrepresentations made publicly.[102] Another is the intentional infliction of emotional distress, which may—but does not always—hinge on false statements.[103] In these instances, akin to defamation, the Court has emphasized the threat to "free and robust debate of public issues", the potential stifling of open dialogue, and the possibility of self-censorship that could arise from awarding damages for false speech. These concerns underscore why the Constitution protects even harmful speech.[104]

The Court has also extended this protection during elections to political candidates' speech, recognizing that the *actual malice* threshold is essential for robust political expression. When reviewing a state law aimed at disqualifying an election winner for making a false statement in good faith and promptly retracting it, the Court cautioned against the "chilling effect [that] such absolute accountability for factual misstatements in the course of political debate" would impose.[105] The Court further observed that, "[i]n a political campaign, a candidate's factual blunder is unlikely to escape the notice of, and correction

---

96 *See* Fed. Trade Comm'n, FTC Coronavirus Warning Letters to Companies, www.ftc.gov/coronavirus/enforcement/warning-letters.

97 *See* Illinois, *ex rel.* Madigan v. Telemarketing Associates, Inc., 538 U.S. 600 (2003).

98 *Alvarez*, 567 U.S. at 719; *see also Alvarez,* 567 U.S. at 734–36 (Breyer, J., concurring); *Alvarez,* 567 U.S. at 747–48 (Alito, J., dissenting).

99 New York Times Co. v. Sullivan, 376 U.S. 254, 271 (1964).

100 *See* Curtis Pub. Co. v. Butts, 388 U.S. 130, 133-34 (1967); Associated Press v. Walker, 398 U.S. 28 (1967).

101 *See* Gertz v. Robert Welsh, Inc., 418 U.S. 323 (1974).

102 *See* Time, Inc. v. Hill, 385 U.S. 374 (1967).

103 *See* Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 56 (1988).

104 *See* Snyder v. Phelps, 562 U.S. 443, 452 (2011) (*quoting* Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749, 760 (1985)).

105 Brown v. Hartlage, 456 U.S. 45, 61 (1982).

by, the erring candidate's political opponent. The preferred First Amendment remedy of 'more speech, not enforced silence,' thus has special force."[106]

When false statements concern public issues, simply requiring that the speech be intentionally false may not satisfy free speech protection. Under strict scrutiny, laws limiting speech must be narrowly tailored to avoid discouraging truthful statements or enabling selective enforcement by government officials based on viewpoint.

Most current justices are concerned that restricting false statements in philosophy, religion, history, and the arts could suppress truthful speech.[107] The Court has justified its broad protection of speech in these areas due to three key factors: (1) the inherent difficulty of defining falsity in such subjective and interpretive fields; (2) the potential for even false speech to contribute meaningfully to public discourse, and (3) the substantial risk that government regulation of false speech could be exploited for political purposes.[108]

II. Online False Speech and Democratic Vulnerabilities

A. *Definition and Nature*

The roots of fake speech, or the intentional spread of false information, can be traced back to ancient history.[109] Even in early civilizations, leaders and influential figures used propaganda and misinformation to sway public opinion and consolidate power.[110]

In Ancient Egypt, pharaohs commissioned monuments and inscriptions that exaggerated their military victories or divine favor to glorify their reigns.[111] Similarly, in Ancient

---

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

107 United States v. Alvarez, 567 U.S. 709, 731-32 (2012)(Breyer, J., concurring).

108 *Id*. at 751-52 (Alito, J., dissenting); *Id*. at 736 (Breyer, J., concurring) (fearing "censorious selectivity"); *Id*. at 723.

109 *See* Phillip M. Taylor, Munitions of the Mind: A History of Propaganda from the Ancient World to the Present 20 (3rd ed. 2003).

110 Taylor elaborates on the use of propaganda and disinformation in early civilizations as follows:

> An early example is the great stela of Eannatum of Lagash (*c*.2550 BC), a round-topped slab depicting Nin-girsu, the god of Lagash, first capturing his enemies in a net and then in a war chariot. On the other side, King Eannatum advances at the head of a well-armed infantry phalanx crushing his enemies underfoot while lions and vultures tear the bodies of the dead. His remaining enemies flee before him and the death sentence is handed out to the defeated king of Umma. Such relics, by their celebratory nature, indicate an awareness of propaganda after the event; standards, decorated shields, and the like demonstrate its use during battle.

Id. at 21.

111 On this topic, Taylor adds the following:

> The gradual shift from war fought in the name of a god to war fought in the name of a king (with the god being reduced to a symbolic presiding influence) may have been due in part to the influence of the Egyptian kings, who developed their own forms of propaganda, in particular spectacular public monuments such as the pyramids and the sphinx. The Pharaohs were among the first to recognize the power of public architecture on a grand scale to demonstrate prestige and dynastic legitimacy. Yet, like the Assyrians, their war propaganda was erratic and sporadic: there was no coherent pattern or organization. Religion was used cynically by rulers to promote loyalty and fear among the ruled. Undoubtedly superstitious themselves, ancient kings backed up their propaganda with terror, both in peace and in war.
>
> *Id.* at 24.

Rome, emperors employed poets and writers to create narratives that portrayed them favorably while downplaying the achievements of their rivals.[112] These early examples of fake speech illustrate how control over information has long been a tool for shaping perceptions and controlling societies.

During the medieval period, fake speech often manifested as religious or political propaganda, as rulers and religious institutions sought to validate their authority. For instance, the Catholic Church circulated false or exaggerated claims to defend its doctrines and discredit competing beliefs or reform movements, such as during the Crusades or the Inquisition.[113] Monarchs also used false claims regarding their divine right to rule or the legitimacy of their lineage to maintain control.[114] These distortions were usually disseminated through speeches, pamphlets, and sermons, where the audience had limited access to contradictory information, allowing false narratives to persist and influence public opinion for centuries.

With the invention of the printing press in the 15th century, fake speech became more accessible to spread, and its reach expanded significantly. During the Protestant Reformation, Catholic and Protestant leaders circulated pamphlets containing exaggerated or false accusations against each other. The proliferation of printed material meant that misinformation could be distributed on a large scale, fueling political and religious conflicts. This represented a turning point in the history of fake speech, as the written word allowed for faster dissemination of false narratives, setting a precedent for the widespread manipulation of information that continues into the modern era.[115]Top of FormBottom of Form

Misinformation is false or inaccurate information shared without the intent to deceive.[116] It encompasses rumors, incorrect facts, or misunderstandings circulated due to a lack of awareness or insufficient verification. For instance, someone might share an outdated statistic or an erroneous interpretation of an event, believing it to be true. Although misinformation is not shared with malicious intent, it can still cause harm by spreading confusion and perpetuating misconceptions, especially when it reaches a wide audience.[117]

---

112 *See Id.* at 35-48.

113 For crusade propaganda, see, e.g., Penny J. Cole, The Preaching of the Crusades to the Holy Land, 1095-1270 (1991); Christoph T. Maier, Preaching the Crusades: Mendicant Friars and the Cross in the Thirteenth Century (1994).

114 *See, e.g.,* Glenn Burgess, *The Divine Right of Kings Reconsidered,* 107 Eng. Hist. Rev. 837 (1992); Mathew Wills, *Making Sense of the Divine Right of Kings*, JSTOR Daily (Dec. 18, 2020), https://daily.jstor.org/making-sense-of-the-divine-right-of-kings/.

115 *See, e.g.,* Andrew Pettegree, *Printing Press and its Impact on the Production, Proliferation, and Readership of Theological Literature*, *in* The Cambridge History of Reformation-Era Theology 9 (Kenneth G. Appold & Nelson H. Minnich eds., Cambridge Univ. Press 2023); Jacob Soll, *The Long and Violent History of Fake News*, POLITICO (Dec. 18, 2016), https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/; Stefan Antripolus, *The Multiplication of Monsters: Misinformation from Gutenberg to QAnon*, Public Books (Mar. 20, 2024), https://www.publicbooks.org/the-multiplication-of-monsters-from-gutenberg-to-qanon/.

116 *See* Fernando Nuñez, *Disinformation Legislation and Freedom of Expression*, 10 U.C. Irvine L. Rev. 783, 785 (2020).

117 *See* Newton Minow & Martha Minow, *Social Media Companies Should Pursue Serious Self-Supervision — Soon: Response to Professors Douek and Kadri,* 136 Harv. L. Rev. F. 428 (2023) (discussing how misinformation on social media platforms can amplify social divisions and risks of violence, highlighting the need for platforms to implement more rigorous self-regulation measures); Daniela C. Manzi, *Managing the Misinformation Market-*

Disinformation, in stark contrast to misinformation, involves deliberately creating and disseminating false information with a clear intent to mislead or manipulate.[118] It is carried out for various purposes, such as political gain, financial profit, or social influence. Unlike misinformation, which is spread unknowingly, disinformation is a calculated effort to deceive.[119] Examples include fake news articles, doctored images, or false narratives strategically planted to sway public opinion or discredit individuals or organizations.

The role of social media algorithms in the rapid spread of misinformation and disinformation cannot be overstated.[120] These algorithms are designed to maximize user engagement by promoting content that elicits strong emotional responses, such as surprise, anger, or fear.[121] As a result, sensational or controversial posts—whether true or false—are more likely to be amplified and widely shared.[122] This creates an environment where misinformation can spread quickly, as users share content without verifying its accuracy, and disinformation can reach large audiences, achieving its deceptive goals more effectively.[123]

In addition, the personalized nature of social media feeds can give rise to *echo chambers*,[124] where users are predominantly exposed to information that aligns with their existing beliefs. This can strengthen biases and make users more susceptible to accepting mis-

---

place: The First Amendment and the Fight Against Fake News, 87 Fordham L. Rev. 2623 (2019) (examining the challenges posed by the rapid spread of misinformation and its impact on democratic processes).

118 *See* Fernando Nuñez, *supra* note 116.

119 *Id.* at 785-86 ("Disinformation is a more serious threat to freedom of expression because it is information that is deliberately created to mislead and influence the public, unlike misinformation, which may be shared under a genuine belief that its contents are truthful.").

120 *See* Haochen Sun, *The Right to Know Social Media Algorithms*, 18 Harv. L. & Pol'y Rev. 1 (2023) (proposing the creation of a "right to know" specifically for social media algorithms).

121 *See* Joshua Benton, *Want to Fight Misinformation? Teach People How Algorithms Work,* Nieman Lab (Sept. 4, 2024), https://www.niemanlab.org/2024/09/want-to-fight-misinformation-teach-people-how-algorithms-work/ (Research indicates that these algorithms prioritize content that sustains engagement, often amplifying information that aligns with users' biases and beliefs. This can result in the proliferation of misinformation that resonates emotionally with users, leading to widespread sharing without verification.).

122 *See* Stephanie Hurder, *The science behind why social media algorithms warp our view of the world*, Fast Company (Aug. 25, 2023), https://www.fastcompany.com/90943919/the-science-behind-why-social-media-algorithms-warp-our-view-of-the-world

(Studies have shown that social media platforms amplify information that sustains engagement, such as prestigious, in-group, moral, and emotional content. This can lead to false polarization, misinformation, and social misperceptions.).

123 *See* William Brady, *Social media algorithms warp how people learn from each other, research shows*, The Conversation (Aug. 21, 2023), https://theconversation.com/social-media-algorithms-warp-how-people-learn-from-each-other-research-shows-211172 (The interaction between human psychology and algorithmic amplification can lead to dysfunction, as social learning supports cooperation and problem-solving. Still, social media algorithms are designed to increase engagement.); Dwi Surjatmodjo et al., *Information Pandemic: A Critical Review of Disinformation Spread on Social Media and Its Implications for State Resilience*, 13 Soc. Sci. 418 (2024) (noting that disinformation spreads six times faster than accurate information, with emotions and platform algorithms playing a significant role in its dissemination); Haochen Sun, *Regulating Algorithmic Disinformation*, 46 Colum. J.L. & Arts 367 (2023) (discussing how recommendation algorithms drive the spread of disinformation on social media networks, highlighting the role of algorithms in amplifying misleading content).

124 *See* Petter Törnberg, *Echo chambers and viral misinformation: Modeling fake news as complex contagion,* PLoS ONE, Sept. 2018, at 1 (footnote omitted) (exploring how echo chambers contribute to the viral spread of misinformation, highlighting the role of social media algorithms in creating ideologically homogeneous communities that facilitate the rapid dissemination of false information).

information and disinformation as truth.[125] The rapid dissemination and extensive reach of social media, in conjunction with these algorithmic inclinations, present challenges in containing the spread of false information and lessening its impact on public discourse and behavior.[126]

B. *Impact on Democratic Processes*

The 2016 U.S. presidential election is a prominent example of how misinformation can undermine trust in democratic institutions, degrade public discourse, and polarize societies. Throughout the presidential campaign, social media platforms such as Facebook and Twitter were inundated with false or misleading information, including fake news articles, manipulated images, and memes. Russian-linked troll farms exerted a significant influence by generating divisive content tailored to specific voter demographics.[127] These tactics magnified societal rifts and nurtured skepticism toward the political process and mainstream media. The proliferation of conspiracy theories, such as those related to Hillary Clinton's emails or Donald Trump's alleged connections to Russia, made it challenging for voters to differentiate reliable information, thereby tarnishing the perceived legitimacy of the election outcome.[128] This distrust persisted beyond the election, contributing to a sustained polarization in American society and diminished confidence in democratic institutions.[129]

The 2016 Brexit referendum exemplifies how misinformation can disrupt public discourse and deepen political divisions. Pro-Brexit campaigns relied on exaggerated claims and outright falsehoods to influence public opinion. For instance, they falsely claimed that leaving the European Union would allow the United Kingdom to redirect £350 million per week to the National Health Service (NHS), a widely discredited figure but still influential among voters.[130] Additionally, misleading narratives about immigration fueled fear and nationalist sentiments, shifting the debate from policy to identity and sovereignty. Following the vote, many Britons expressed regret, citing being misled by false information.[131] The misinformation influenced the vote's outcome and prolonged societal division, eroding public trust in the media and government.[132]

---

125 *Id.*

126 *Id.* at 17-22.

127 Jen Kirby, *What to know about the Russian troll factory listed in Mueller's indictment*, Vox (Feb. 16, 2018), https://www.vox.com/2018/2/16/17020974/mueller-indictment-internet-research-agency.

128 Darren Samuelsohn, *The most dangerous conspiracy theory of 2016,* POLITICO (Sept. 28, 2016), https://www.politico.com/story/2016/09/2016-election-conspiracy-theory-rigged-228807.

129 *See, e.g.,* Hunt Allcott & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election,* 31 J. Econ. Persp. 211 (2017).

130 *See £350 million EU claim "a clear misuse of official statistics"*, Full Fact (Sept. 19, 2017), https://fullfact.org/europe/350-million-week-boris-johnson-statistics-authority-misuse/.

131 *See Final Say: The misinformation that was told about Brexit during and after the referendum,* The Independent (Jul. 28, 2018), https://www.independent.co.uk/news/uk/politics/final-say-brexit-referendum-lies-boris-johnson-leave-campaign-remain-a8466751.html.

132 Uta Staiger, *Brexit: five years after the referendum, here are five things we've learned*, The conversation (Jun. 23, 2021), https://theconversation.com/brexit-five-years-after-the-referendum-here-are-five-things-weve-learned-162974.

The COVID-19 pandemic underscored the perils of misinformation in a global health crisis. From the virus's origins to vaccine efficacy, conspiracy theories and false information spread across social media and alternative news sources. Some of the most damaging narratives included claims that COVID-19 was a hoax, that vaccines contained microchips, and that unproven treatments were effective.[133] These false claims undermined trust in public health institutions and led to significant public resistance to scientifically recommended measures such as mask mandates, social distancing, and vaccinations.[134] The polarized responses to these measures often aligned with political affiliations, further dividing societies and undermining collective efforts to control the virus's spread. In this case, misinformation did not just impact public discourse; it had real, life-or-death consequences, underscoring the destructive potential of false narratives on societal cohesion and institutional trust.[135]

These case studies show that misinformation is not merely a matter of factual inaccuracy —it actively erodes democratic norms, undermines public trust, and deepens societal divides. The 2016 U.S. election, Brexit, and the COVID-19 pandemic illustrate the urgent need for strategies to counter misinformation and restore faith in democratic processes, public health systems, and media institutions. Without effective countermeasures, the continued spread of false information threatens to weaken the very foundations of democratic societies.

C. *Legal and Ethical Tensions*

Balancing the need to combat misinformation while protecting free speech is complex and contentious. Misinformation can have serious consequences, such as undermining public health efforts, influencing elections, and spreading harmful stereotypes.[136] To address these risks, platforms and governments may implement measures like content moderation, fact-checking, and regulatory policies to limit the spread of false information.[137] However, these interventions can raise concerns about overreach and censorship, potentially stifling legitimate expression and infringing on individuals' rights to share their opinions and access diverse viewpoints.[138]

---

133 Jack Goodman & Flora Carmichael, *Coronavirus: Bill Gates 'microchip' conspiracy theory and other vaccine claims fact-checked*, BBC (May 19, 2020), https://www.bbc.com/news/52847648.

134 Alistair Coleman, '*Hundreds dead' because of Covid-19 misinformation*, BBC (Aug. 12, 2020), https://www.bbc.com/news/world-53755067.

135 *See* Sarah Gibbens, *A guide to overcoming COVID-19 misinformation,* National Geographic (Oct. 22, 2020) https://www.nationalgeographic.com/science/article/guide-to-overcoming-coronavirus-misinformation-infodemic; *Debunking COVID-19 myths*, Mayo Clinic (Sept. 7, 2024), https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-myths/art-20485720.

136 Stephan Lewandowsky et. al., *Disinformation Is the Real Threat to Democracy and Public Health*, Scientific American (Jan. 30, 2024), https://www.scientificamerican.com/article/disinformation-is-the-real-threat-to-democracy-and-public-health/.

137 *See* Darrell M. West, *How to combat fake news and disinformation*, Brookings (Dec. 18, 2017), https://www.brookings.edu/articles/how-to-combat-fake-news-and-disinformation/.

138 *Moderating online content: fighting harm or silencing dissent?,* United Nations (Jul. 23, 2021), https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent.

Content moderation involves removing or restricting content deemed misleading or harmful. However, determining what constitutes misinformation can be subjective and challenging, and this process risks silencing voices that may offer alternative perspectives or legitimate critiques. Especially if moderation guidelines are applied inconsistently or influenced by political or social biases, individuals may perceive such measures as an infringement on their freedom of speech.[139] This can fuel distrust in platforms and regulatory authorities, potentially driving the conversation to less regulated, more opaque channels.[140]

Fact-checking is a valuable tool in combating misinformation by providing accurate information to counter false claims. While less intrusive than content removal, it can still lead to tensions. Critics argue that fact-checkers may not always be impartial, and labeling or flagging content as false can be perceived as a form of bias or suppression.[141] Additionally, users whose posts are flagged or debunked might feel unjustly targeted, further entrenching their beliefs and contributing to a polarized environment where free discourse is hampered by mutual suspicion.[142]

Regulation adds another layer of complexity, as laws to curb misinformation must be carefully crafted to avoid infringing on free speech rights. Broad or vague regulations can be misused to target dissent or censor unpopular opinions under the guise of combating misinformation.[143] This creates a delicate balance, as it's necessary to prevent the harm caused by misinformation while ensuring that such measures do not erode the foundational principles of free expression and democratic participation.[144] Navigating this tension requires a nuanced approach that upholds the integrity of information without compromising the essential freedoms that support an open society.[145]

---

139 *See* Christopher Keleher, *The Antidote of Free Speech: Censorship During the Pandemic*, 73 CATH. U. L. REV. 213 (2024) (discusses the tension between free speech and government intervention during the pandemic, highlighting the conflict between promoting public health measures and maintaining the right to free expression); *See also* Brenda Dvoskin, *The Thorny Problem of Content Moderation and Bias*, CENTER FOR DEMOCRACY AND TECHNOLOGY (Jul. 3, 2019), https://cdt.org/insights/the-thorny-problem-of-content-moderation-and-bias/.

140 *See* Joan Donovan, *Why social media can't keep moderating content in the shadows*, MIT TECHNOLOGY REVIEW (Nov. 6, 2020), https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/.

141 *See* MARIAVITTORIA MASOTINA ET AL., IMPARTIALITY AND COGNITIVE BIAS IN THE FACT-CHECKING PROCESS: AN OVERVIEW (2023), https://datalab.luiss.it/ricerche/fact-checking-process/.

142 Andrea Carson et al., *Fact-checking can actually harm trust in media: new research*, THE CONVERSATION (Mar. 3, 2022), https://theconversation.com/fact-checking-can-actually-harm-trust-in-media-new-research-176032; Jerusalem Demsas, *When Facts-Checks Backfire*, THE ATLANTIC (Sept. 17, 2024), https://www.theatlantic.com/podcasts/archive/2024/09/fact-check-polarization-opinion-politics/679898/.

143 Rasmus Kleis Nielsen, *How to respond to disinformation while protecting free speech*, REUTERS INSTITUTE (Feb. 19, 2021), https://reutersinstitute.politics.ox.ac.uk/news/how-respond-disinformation-while-protecting-free-speech.

144 *See, e.g.,* Leslie Gielow Jacobs, *Freedom of Speech and Regulation of Fake News,* 70 AM. J. COMP. L. i278 (2022), https://doi.org/10.1093/ajcl/avac010.

145 *See* Melissa De Witte, *What Stanford research reveals about disinformation and how to address it*, STANFORD REPORT (Apr. 13, 2022), https://news.stanford.edu/stories/2022/04/know-disinformation-address.

III. Content Moderation and Legal Frameworks

A. *Content Moderation Practices*

Major social media platforms like Facebook, Twitter (now X), and YouTube have developed complex strategies for moderating harmful content. These strategies aim to identify, label, or remove content that violates community guidelines, including hate speech, misinformation, and graphic violence.[146] The platforms use AI-based tools and human moderation to balance rapid responses with contextual understanding. While these efforts aim to maintain a safe online environment, they face challenges and criticisms.

AI plays a central role in moderating harmful content due to its ability to operate at scale and speed. AI algorithms detect content that violates policies, such as hate speech, spam, or explicit material, by analyzing text, images, and video metadata.[147] For instance, Facebook employs AI to pre-screen posts, flag potential violations, and prioritize them for review.[148] YouTube's AI, known as *Content ID*, can identify copyrighted and inappropriate content by scanning billions of videos. Twitter also uses AI to detect abusive language or manipulated media, applying filters to remove content or issue user warnings.

However, AI-driven moderation is not foolproof. It struggles with nuanced language, cultural differences, and context, often resulting in false positives or negatives.[149] For example, AI may have difficulty distinguishing between hate speech and satire, or it might fail to catch subtle misinformation because it cannot evaluate the broader context of a post. Additionally, AI can be biased, reflecting the limitations of the data it is trained on, which may lead to disproportionate content removal affecting marginalized communities.[150]

Despite the reliance on AI, human moderators are essential to content moderation on these platforms. Human moderators intervene when AI tools encounter ambiguous content that requires more profound understanding, such as evaluating political speech, satire, or context-specific humor. For instance, YouTube employs moderators to review flagged videos that may contain sensitive or borderline content.[151] Facebook uses human

---

146 *See* Niva Elkin-Koren et al., *Social Media as Contractual Networks: A Bottom-Up Check on Content Moderation*, 107 Iowa L. Rev. 987 (2022) (suggesting that contractual network theory can create a *bottom-up check* on content moderation practices, empowering users to challenge unfair moderation decisions and promote a more pluralistic public discourse).

147 *See* Theodore F. Claypoole, *Social Media: Modern Content Moderation(?)*, Nat'l L. Rev. (Jan. 19, 2021), https://natlawreview.com/article/new-age-content-moderation.

148 James Vincent, *Facebook Says Its AI Moderator Can Enforce Rules More Precisely*, The Verge (Nov. 13, 2020), https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation.

149 Merlyna Lim & Ghadah Alrasheed, *Beyond a technical bug: Biased algorithms and moderation are censoring activists on social media,* The Conversation (May. 16, 2021), https://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669.

150 *See* Gabriel Nicholas & Aliya Bhatia*, The Dire Defect of 'Multilingual' AI Content Moderation*, Wired (May. 23, 2023), https://www.wired.com/story/content-moderation-language-artificial-intelligence/; Carol Anderson, *Guardrails on Large Language Models, Part 4: Content Moderati*on, AVID (Apr. 11, 2023), https://avidml.org/blog/llm-guardrails-4/.

151 James Vincent, *YouTube brings back more human moderators after AI systems over-censor*, The Verge (Sept. 21, 2020), https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns.

reviewers for content that are reported by users or flagged by AI, focusing on misinformation, hate speech, and violent imagery.[152]

While human moderation offers a more nuanced approach, it also faces limitations. Reviewers can experience mental health challenges due to the nature of the content they examine, leading to potential burnout and reduced accuracy.[153] Moreover, the vast amount of content on social media means that human reviewers cannot catch everything promptly, making AI a necessary, though imperfect, partner. Facebook highlights this reality, reporting that, in a single quarter, it acted on 22.1 million pieces of hate speech content, underscoring the scale of the challenge of content moderation.[154]

The hybrid AI and human moderation model reflects the complexity of regulating online content at scale. As platforms encounter increasing regulatory pressures and public scrutiny, they refine their moderation strategies, investing in better AI tools and more effective human oversight. Despite these efforts, finding the right balance remains difficult. Platforms must navigate the challenge of upholding free speech while preventing harm and addressing transparency, accountability, and bias concerns. The future of content moderation will likely involve more sophisticated AI tools that can better understand context, combined with increased transparency about moderation decisions to foster user trust.

B. *Legal Standards and Challenges*

Section 230 of the *Communications Decency Act*, enacted in 1996, is a cornerstone of U.S. online legal standards. It provides extensive liability immunity to internet service providers and social media companies, shielding them from being held accountable for user-generated content.[155] Specifically, section 230(c)(1) asserts that platforms should not be regarded as the "publisher or speaker" of content produced by third parties.[156] This legal immunity empowers platforms to host user-generated content with reduced risk of litigation concerning defamation, misinformation, or other unlawful material.

Furthermore, section 230 encourages platforms to engage in *good faith* content moderation, permitting them to remove or curtail content considered objectionable regardless

---

152 Barbara Ortutay, *Facebook's system approved dehumanizing hate speech*, PBS News (Jun. 9, 2022), https://www.pbs.org/newshour/world/facebooks-system-approved-dehumanizing-hate-speech.

153 Jennifer Beckett, *We need to talk about the mental health of content moderators,* The Conversation, (Sept. 27, 2018), https://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830; *ICMEC Creates Framework To Protect Human Social Media Content Moderators From Psychological Harm*, MENAFN (Nov. 1, 2024), https://menafn.com/1108842597/ICMEC-Creates-Framework-To-Protect-Human-Social-Media-Content-Moderators-From-Psychological-Harm.

154 Elizabeth Culliford & Katie Paul, *Facebook offer up first-ever estimate of hate speech prevalence on its platform,* Reuters (Nov. 19, 2020), https://www.reuters.com/article/technology/facebook-offers-up-first-ever-estimate-of-hate-speech-prevalence-on-its-platform-idUSKBN27Z2QX/; *see also* Theara Coleman*, Pros and cons of social media content moderation,* THE WEEK (Oct. 9, 2023), https://theweek.com/tech/pros-and-cons-of-social-media-content-moderation.

155 *See Section 230 as First Amendment Rule*, 131 Harv. L. Rev. 2027 (2018) ("Section 230 of the *Communications Decency Act* of 1996 has been lauded as 'the most important law protecting internet speech' and called 'perhaps the most influential law to protect the kind of innovation that has allowed the Internet to thrive.'").

156 47 U.S.C.A. § 230(c)(1) (1996).

of its legal standing. This provision fosters proactive content management, allowing platforms to enforce community guidelines without fearing being penalized for censorship. Nonetheless, section 230 has come under scrutiny, with critics claiming that it excessively shields platforms, allowing the proliferation of detrimental content.[157] Recent proposals to reform section 230 aim to enhance platform accountability concerning the content that incites violence, spreads falsehoods, or contravenes other legal frameworks. However, these reforms must balance the need for free speech and innovation.

In contrast, the *European Union's Digital Services Act* (hereinafter, "DSA"), adopted in 2022, presents a comprehensive regulatory framework tailored to ensure accountability among digital platforms, particularly concerning illegal content and services.[158] While Section 230 emphasizes platform immunity, the DSA prioritizes accountability, especially for large platforms with over 45 million users in the European Union.[159] It mandates proactive measures, including transparent content moderation processes, risk assessments, and deploying algorithms to identify and mitigate harmful or unlawful content dissemination.

The DSA imposes more rigorous obligations for content removal, necessitating swift action on flagged illegal material while preserving user rights through established appeal mechanisms.[160] Additionally, it enforces transparency in algorithmic decision-making and advertising practices to curb misinformation and harmful content dissemination. By placing explicit responsibilities on platforms to avert the spread of illegal material, the DSA narrows the scope of immunity compared to section 230. Nevertheless, it aims to safeguard freedom of expression by requiring proportionate actions to prevent excessive censorship. This reflects the European Union's commitment to balancing public safety, rights of expression, and the demand for transparency and accountability.

---

157 Kira M. Geary, *Section 230 of the Communications Decency Act, Product Liability, and a Proposal for Preventing Dating-App Harassment*, 125 Penn St. L. Rev. 501 (2021); Heidi Tripp, *All Sex Workers Deserve Protection: How FOSTA/SESTA Overlooks Consensual Sex Workers in an Attempt to Protect Sex Trafficking Victims*, 124 Penn St. L. Rev. 219 (2019); Matthew G. Jeweler, *The Communications Decency Act of 1996: Why § 230 Is Outdated and Publisher Liability for Defamation Should Be Reinstated Against Internet Service Providers*, 8 U. Pitt. J. Tech. L. & Pol'y 3 (2008); Brandon Salter & Dhillon Ramkhelawan, *Section 230 Immunity: How the Trump Era Has Exposed the Current Conflict Between the First Amendment and the Good Samaritan Clause in the Modern Public Square*, 43 U. Ark. Little Rock L. Rev. 239 (2020).

158 *See, e.g.,* Ioanna Tourkochoriti, *The Digital Services Act and the EU as the Global Regulator of the Internet,* 24 Chi. J. Int'l L. 129 (2023).

159 *See* Kimberly A. Fry, *The Fate of Section 230*, 22 Colo. Tech. L.J. 361 (2024) (proposing that the U.S. could model its reforms after the *European Union's Digital Services Act* (DSA), which focuses on platform accountability, transparency in moderation practices, and protection of user rights).

160 Fry adds the following in her article:

> With the passage of the DSA, online platforms went from 'effectively no regulation to heavy regulation.' Before the DSA, Google could get away with saying the company was offering more 'visibility' into content moderation decisions and different ways for users to contact the company, without offering specifics. Under the DSA, Google and other platforms became legally obligated to provide more information behind why posts are removed, among other heavy regulations. To Big Tech, adapting to the DSA's new rules meant survival, at least in the European market. These titans of the Internet industry rushed to meet their compliance obligations with the EU by rolling out 'new ways' for European users to flag illegal online content and dodgy products" and "quickly and objectively" removing this flagged content from their platforms.

Id. at 396-97 (footnotes omitted).

Other nations have instituted similar frameworks globally to address platform liability and content moderation.[161] *Germany's Network Enforcement Act* (NetzDG), enacted in 2017, requires platforms to expeditiously remove *manifestly unlawful* content within twenty-four hours or face significant penalties.[162] This legislation has garnered criticism for fostering over-compliance and stifling legitimate discourse, yet it has set a precedent influencing subsequent European regulations like the DSA.

In India, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules of 2021 mandate that platforms act promptly against illegal content, particularly when flagged by government authorities, and require the appointment of local compliance officers to enhance accountability.[163] Conversely, *Australia's Online Safety Act* (2021) grants the government the authority to issue *takedown notices* for harmful content, emphasizing cyberbullying and abusive behavior, and holding platforms accountable for non-compliance.[164]

Legal frameworks such as section 230, the DSA, and regional regulations differ in their approaches to platform liability, content moderation, and the balance of protection, accountability, and transparency. While section 230 encourages innovation through broad immunity, the DSA and analogous laws adopt a more interventionist paradigm, emphasizing user safety and proactive content management. The global trend indicates a shift toward increased platform accountability, accompanied by ongoing discourse regarding the ramifications of free speech, overreach, and technological innovation. As platforms continue influencing public discourse, these legal frameworks are pivotal in delineating platform responsibility and user rights in the evolving digital landscape.

C. *Censorship vs. Free Speech*

Regulating harmful online content is crucial for safeguarding public safety, combating misinformation, and maintaining community standards. However, this initiative often treads a precarious line between addressing harmful content and enabling censorship, particularly when regulations lack specificity or are too expansive. While the primary objective of these laws is to mitigate issues like violence, hate speech, and misinformation—especially those that jeopardize public health—poorly designed regulations can be weaponized to suppress dissent and limit free expression. This risk is heightened in authoritarian regimes, where such measures are frequently leveraged to target political rivals, silence independent media, or curtail public criticism.

In authoritarian contexts, laws intended to curb harmful content frequently transform into instruments of control and repression. For instance, legislation against *fake news* has been exploited in nations such as Russia, Turkey, and China to justify harsh crackdowns

---

161 *See, e.g.,* Patrick Zurth, *The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability,* 31 Fordham Intell. Prop. Media & Ent. L.J. 1084 (2021).

162 *See* Emily Lagg, *Stormy Waters for the Internet's Safe Harbor: The Future of Section 230,* 71 Rutgers U. L. Rev. 763, 784 (2019).

163 Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (India).

164 *Online Safety Act 2021* (Cth) (Austl.).

on political adversaries, activists, and journalists. In Russia, the government has employed disinformation laws to restrict access to independent news sources and social media platforms that challenge state narratives, especially during pivotal events like the invasion of Ukraine.[165] Similarly, in China, extensive censorship is rationalized under the guise of *public safety* or *harm prevention*, yet fundamentally serves to consolidate government control over information and suppress dissent regarding the Chinese Communist Party.[166]

In April 2020, during the COVID-19 pandemic, Puerto Rico's Governor Wanda Vázquez signed an amendment to the *Puerto Rico Department of Public Safety Act*.[167] The amendment included measures to prevent the spread of misinformation related to government proclamations or executive orders concerning COVID-19 and other disasters. The law made it illegal for media outlets or social media accounts to disseminate *false information* related to such proclamations or orders, with penalties including imprisonment and fines.[168]

The enactment of this law raised concerns about freedom of speech and the press. In May 2020, the American Civil Liberties Union (ACLU), representing journalists Sandra Rodríguez Cotto and Rafelli González Cotto, filed a lawsuit challenging the constitutionality of the law.[169] They argued that the law's broad and ambiguous definitions of *false information* could be used to suppress legitimate reporting and criticism of the government's handling of emergencies, thus violating First Amendment rights.[170]

In April 2023, the U.S. District Court for the District of Puerto Rico ruled in favor of the plaintiffs, declaring the law unconstitutional.[171] The court explained that the law's broad scope posed a risk of partisan abuse or selective enforcement, allowing the government to suppress or stifle speech that contradicted its official narrative.[172] The court emphasized

---

165 *Russia's Crackdown on Independent Media and Access to Information Online*, Center for Strategic & International Studies (Mar. 30, 2022), https://www.csis.org/analysis/russias-crackdown-independent-media-and-access-information-online; Victor Jack, *Russia expands laws criminalizing 'fake news'*, POLITICO (Mar. 22, 2022), https://www.politico.eu/article/russia-expand-laws-criminalize-fake-news/.

166 Yaqiu Wang, *Censorship Is No Solution to China's Public Safety Problem*, The Diplomat (Jun. 28, 2024), https://thediplomat.com/2024/06/censorship-is-no-solution-to-chinas-public-safety-problem/.

167 Amend Section 6.14 of Act No. 20-2017, Act No. 35-2020, 2020 LPR 200 (currently in Puerto Rico Department of Public Safety Act, Act No. 20-2017, 25 LPRA § 3654 (2024)).

168 *New Puerto Rico law threatens jail time for spreading 'false information' about COVID-19*, Committee to Protect Journalists (Apr. 8, 2020), https://cpj.org/2020/04/new-puerto-rico-law-threatens-jail-time-for-spread/.

169 *ACLU Challenges Puerto Rico COVID-19 'Fake News' Laws*, ACLU (May. 20, 2020), https://www.aclu.org/press-releases/aclu-challenges-puerto-rico-covid-19-fake-news-laws.

170 The ACLU press release states the following:

> The ACLU argues that Puerto Rico's *fake news* laws, which do not require the government to demonstrate that the speaker knew the speech was false, violate the First and Fourteenth Amendments because of their vague terminology and broad sweep. The laws give people far too little guidance on what speech may constitute a crime and government far too much discretion in deciding whom to prosecute. As a result, the laws chill a great deal of reporting on the COVID-19 crisis and other emergencies, because journalists risk prosecution if the government disputes the accuracy of their reporting.

Id.

171 Rodríguez-Cotto v. Pierluisi-Urrutia, 668 F. Supp.3d 77, 110-11 (D.P.R. 2023).

172 *Id.* at 105 ("Much like the *Stolen Valor Act*, this broad sweep is at odds with the First Amendment. Besides, it fails as being impermissibly underinclusive.").

the crucial role of free speech during crises, stating, "[t]he watchdog function of speech is never more vital than during a large-scale crisis."[173]

Even in democratic contexts, applying necessary regulations might inadvertently suppress legitimate discourse due to broad interpretations. For example, *Germany's Network Enforcement Act* (NetzDG), which mandates the quick removal of illegal content, has faced criticism for prompting platforms to engage in over-compliance, thereby stifling lawful yet controversial speech.[174]

Achieving a balance between effective regulation and preserving free expression necessitates several safeguards. Transparent, precisely defined laws with explicit criteria for identifying *harmful content* can mitigate the risk of regulatory misuse. Integrating independent oversight mechanisms and judicial review processes is vital to ensure that content moderation decisions are equitable and devoid of political bias. Additionally, mandating transparent appeal processes for users can provide necessary recourse against wrongful content removal.

Striking this balance is particularly challenging in a rapidly evolving digital landscape where information circulates across borders. This complexity is further exacerbated by the political, cultural, and economic contexts influencing various countries' regulatory approaches. While democratic nations aim to uphold individual rights, authoritarian regimes often disguise acts of censorship as regulatory measures. Therefore, global initiatives establishing best practices for content moderation must consider these dynamics, prioritizing transparency, accountability, and due process to protect safety and fundamental liberties.

The interplay of regulation and censorship significantly impacts the essence of free societies. Overly restrictive rules can erode the open discourse essential to democratic governance. Conversely, ineffective regulation may allow harmful content to increase, inciting violence, spreading misinformation, and undermining public trust. As digital platforms increasingly influence public discourse, the global community faces an urgent imperative: cultivating an online environment that is safe and conducive to freedom, resisting authoritarian abuse while promoting accountability.

IV. Legal and Policy Proposals for the Future

A. *Regulatory Approaches*

One of the most discussed proposals for regulating social media companies is increasing transparency in algorithms. Social media algorithms determine which content is promoted, downranked, or hidden, profoundly shaping users' experiences and perceptions. Proposals for algorithmic transparency generally call for companies to disclose how their algorithms work, the factors they consider, and how they impact content visibility. Greater transparency could help users understand why they see specific posts, identify potential

---

173 *Id.* at 110.

174 *Germany: Flawed Social Media Law*, Human Rights Watch (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

biases, and encourage platforms to be more accountable for the content they amplify.

However, there are challenges to implementing algorithmic transparency. Social media companies often argue that disclosing too much algorithm information could enable malicious actors to game the system or compromise proprietary technology. There is also a risk of information overload (providing highly technical details about algorithms may be too complex for most users to understand, limiting the impact of transparency efforts). Despite these concerns, many policymakers believe that increasing transparency is essential to fostering trust, and proposals often require clear explanations for content decisions or independent audits of algorithms by third parties.

Another regulatory proposal focuses on strengthening fact-checking requirements for social media companies. Platforms like Facebook and X already collaborate with third-party fact-checkers to label misleading or false information. Proposals to mandate fact-checking seek to expand these efforts, ensuring that platforms consistently identify and flag false information, especially regarding critical issues like elections, public health, and national security.

While this approach can limit the spread of misinformation, it also raises questions about the feasibility and potential biases of fact-checking. Fact-checking is often subjective, especially when dealing with complex or controversial topics, making it difficult to establish universally accepted criteria. Additionally, implementing fact-checking at scale is challenging, given the sheer volume of content generated daily. Critics also argue that mandated fact-checking could lead to overreach, with platforms potentially acting as *arbiters of truth* and stifling legitimate debate. Balancing rigorous fact-checking with free speech protections remains a critical challenge in designing these proposals.

Changes to intermediary liability laws, such as the U.S. section 230 of the *Communications Decency Act* or similar laws in other countries, represent a more structural approach to regulating social media companies. Reforms typically aim to hold platforms more accountable for illegal content, such as hate speech, harassment, or misinformation, by reducing their immunity from liability. Proponents argue that these changes would incentivize platforms to improve moderation practices, invest in better tools, and take proactive measures to prevent harm.

However, altering intermediary liability laws could have unintended consequences. Platforms might adopt stricter moderation policies to minimize legal risks, potentially leading to over-censorship and limiting free expression. Smaller platforms could struggle to comply with heightened liability requirements, consolidating power among more prominent players with the resources to manage increased regulatory burdens. To address these concerns, proposals often include tiered liability systems, providing smaller platforms with more lenient standards and allowing for different levels of regulation based on platform size and reach.

While these proposals for regulating social media companies aim to enhance transparency, reduce misinformation, and increase accountability, they must balance these goals with protecting free expression and fostering innovation. Requiring transparency and fact-checking can encourage responsible content sharing. Still, ensuring these measures do not hinder users' ability to engage in open debate or share diverse viewpoints is vital. Similarly, liability laws should be adjusted carefully to prevent stifling new platforms

or deterring legitimate speech. Moving forward, any regulatory approach will require a nuanced, adaptable strategy that addresses the dynamic nature of social media while safeguarding fundamental rights.

B. *Protecting Democratic Speech*

Designing legal frameworks that balance free speech protections with measures to combat misinformation is complex but essential. The objective is to create regulations that address harmful misinformation without suppressing legitimate dissent or silencing minority voices. A well-balanced framework should emphasize transparency, accountability, and proportionality, ensuring efforts to counter misinformation do not infringe upon robust free speech rights.

Transparency should be fundamental in any legal framework designed to combat misinformation. Laws could require social media platforms to disclose moderation policies, algorithms, and decision-making criteria. Being transparent about how content is flagged, removed, or downranked helps build public trust and allows users to understand why certain content is promoted or restricted. To ensure this principle does not disproportionately impact minority voices or dissenting opinions, regulations could mandate independent audits of algorithms, focusing specifically on bias against underrepresented groups.

Additionally, platforms should be mandated to offer clear explanations when content is flagged or removed, accompanied by accessible appeal mechanisms. This approach allows users to contest decisions, promoting an environment where diverse perspectives can still be heard. By enforcing transparency, platforms can be held accountable for their moderation practices, which reduces the risk of overreach or discriminatory enforcement.

Legal frameworks should include targeted and proportionate measures to prevent the spread of misinformation without suppressing dissent. Laws should differentiate between types of misinformation based on their potential harm. For instance, harmful content such as election interference, public health misinformation, or incitement to violence should be addressed more rigorously. In contrast, less toxic or debatable content can be approached with a lighter touch. This proportionality ensures that regulations do not treat all misinformation equally and do not suppress legitimate criticism or satire.

For example, platforms could be required to implement tiered moderation systems that vary based on the severity of the content. While severe misinformation—such as false medical advice or election-related falsehoods—may warrant swift removal, less harmful content could be subject to labeling, demotion, or fact-checking rather than deletion. This approach prevents immediate harm and preserves space for diverse discourse and criticism.

Legal frameworks should include safeguards against discriminatory moderation practices to protect minority voices and dissenting opinions. Independent oversight bodies could be established to ensure that content moderation does not disproportionately target specific groups or perspectives. These bodies should feature diverse representation to account for various cultural contexts, guaranteeing that moderation is applied relatively across different communities.

Additionally, regulations could include protections for anonymous speech, which has historically been vital for marginalized communities and political dissenters. While ano-

nymity can contribute to spreading misinformation, it is also crucial for free expression in repressive environments. Striking a balance between these competing interests would involve allowing anonymous speech while requiring platforms to create robust tools to detect coordinated disinformation campaigns or harmful activities without compromising user privacy.

To complement these legal frameworks, governments should invest in media literacy and public awareness campaigns that empower users to identify misinformation independently. By focusing on education and critical thinking skills, society can develop a more resilient public that is less reliant on legal measures to distinguish fact from falsehood. Legal frameworks prioritizing media literacy encourage citizens to engage critically with content while promoting a culture of dialogue and debate.

Legal frameworks that balance free speech and misinformation countermeasures require a nuanced, multi-faceted approach. Principles such as transparency, proportionality, protection of minority voices, and media literacy can guide these regulations to address harmful misinformation without suppressing diverse perspectives or dissent. These frameworks should promote a healthier information ecosystem that fosters free expression and accountability.

Conclusion

The regulatory challenges surrounding free speech on social media underscore the intricate balance between safeguarding democratic values and mitigating the impact of misinformation. In contemporary democratic societies, free speech is foundational; however, the digital landscape has magnified its advantages and pitfalls. Social media platforms have revolutionized public discourse by enhancing communicative accessibility, but this expansive openness simultaneously facilitates the rapid dissemination of misinformation, hate speech, and other detrimental content.

The dynamics of digital communication reveal that the unchecked proliferation of false information can erode public trust, exacerbate societal polarization, and undermine democratic processes. Effective regulatory frameworks must navigate the delicate tension between harm prevention and the protection of free expression. Strategies such as content moderation, fact-checking, and regulatory oversight are pivotal, yet they necessitate implementation characterized by transparency, accountability, and a commitment to preserving free speech.

Ultimately, the integrity of free speech in the digital realm demands legal innovation and a comprehensive commitment to cultivating an informed, critical, and resilient citizenry. As social media continues to evolve, our strategies for preserving the free exchange of ideas must also advance, ensuring that the digital public square remains safe and genuinely open.