SPEECH CONTESTATION BY DESIGN: DEMOCRATIZING SPEECH GOVERNANCE BY AI

NIVA ELKIN-KOREN* AND MAAYAN PEREL**

ABSTRACT

The online elaboration of speech norms is enduring a decisive transformation, threatening the vital prospects of democratic contestation, which enable democracies to thrive. In this Article, we demonstrate how a critical space for social deliberation and negotiation of the desirable boundaries of free speech is "lost in translation" as we shift from governance by law to governance by Artificial Intelligence (AI).

The configuration of AI speech filtering systems facilitates a frictionless flow of information—a signature trait of the digital economy, and of social media in particular. It is driven by a probabilistic decisionmaking process based on formal definitions and optimization dynamics, which are designed to enable speedy detection of harmful content. AI speech moderation systems effectively formulate data-driven decision rules, which reflect a single, pre-defined and potentially biased tradeoff. It currently lacks, however, adequate contesting mechanisms and fails to facilitate the vital normative space necessary for deliberating the disagreements in society regarding the scope of free speech.

In contrast, governance of online speech by law is discursive, permitting different tradeoffs to coexist. Speech governance by law further facilitates a shared ground for voicing dissent and addressing it. By its institutional design, and various procedures and practices, governance by law in liberal democracies facilitates democratic contestation, and it is therefore better equipped to sustain divided societies in the absence of deeper normative consensus.

The absence of democratic contestation in speech governance by AI undermines the legitimacy of speech norms, precludes public engagement in checking and testing which values are embedded in

* Professor, Tel-Aviv University, Faculty of Law; Faculty Associate, Berkman Klein Center at Harvard University.

** Assistant Professor, Netanya Academic College, Faculty of Law.

This research was supported by the Israel Science Foundation (grant No. 1820/17). We thank Elettra Bietti, Michael Birnhack, Julie Cohen, Ellen Goodman, Robert Post, Amnon Reichman, and Eli Salzberger for their excellent feedback. We further thank the participants of the 2021 Digital Governance in the Times of Covid-19 workshop, the participants of the 2021 Data Law and Ethics Research Workshop, the Tel Aviv Faculty Colloquium, the Georgetown Law School Technology Law and Policy Colloquium, Politicizing the Digital Medium Workshop, Wissenschaftskolleg zu Berlin June 2022, and the Freedom of Expression Scholars Conference 10 (2022).

algorithmic tradeoffs, and interferes with the pluralistic aspiration to develop social norms through democratic processes of public engagement and deliberation.

This Article proposes to introduce speech contestation by design in order to legitimize the way AI systems currently shape online speech norms. Inspired by the contestation mechanisms of the law, such as separation of powers and adversarial legal procedures, this Article suggests separation of functions and contesting algorithms as exemplary design features of AI systems of speech governance. Embedding such design features into AI systems of speech moderation may enable ongoing social dialogue between diversified views regarding the limits of free speech. Legal policy pertaining to automated speech moderation by digital platforms should therefore focus on promoting such design interventions.

INTRODUCTION

The online elaboration of speech norms is enduring a decisive transformation. The digital public sphere is mediated by digital platforms deploying Artificial Intelligence (AI) and Machine Learning (ML) to moderate online speech. Consequently, the norms that govern online discourse are generated automatically by non-transparent algorithms, which are driven by data. As we shift from governing speech by law and legal institutions to speech governance by AI, a critical space for contesting the desirable boundaries of free speech is "lost in translation."

The current design of AI systems, which governs online speech, leaves little room for social participation in deliberating, negotiating, and collectively deciding the scope of free speech. Yet, sustaining a discursive social dialogue between diverse values and opinions is a key feature of liberal democracies, enabling disagreement while at the same time keeping society whole. This is especially critical in contemporary times of major social and political transitions, where the scope of free speech in liberal democracies is called into question.[1] Is it possible to sustain democratic contestation in speech governance by AI?

Consider, for instance, the case of Manny Marotta, a history graduate from the University of Pittsburgh. Marotta has created Instagram and Twitter accounts, named *100 Years Ago Live*, to describe history in the language of modern social tools.[2] On July 29, 2021, Marotta posted a short news report look-alike post on his *1921 Live @100YearsAgoLive* account, reporting the election of Adolf Hitler as the new leader of the National Socialist German Workers' Party. The post included a black-and-white photo of Hitler as part of Marotta's attempt to put "readers in the mood of the era" while at the same time keeping their thoughts in the present.[3] Instagram automatically removed the post for violating its "community guidelines" and further rejected Marrota's appeal, confirming that his post related to "violence or dangerous organizations."[4] Instagram's algorithmic speech moderation system, like other systems deployed by social media platforms,[5] purports to combat unwarranted content, such as hate speech, violent extremism, terrorism, conspiracy theories,

1. *See infra* Section II.C.

2. Matt Taibbi, *Meet the Censored: Hitler*, RACKET NEWS (July 30, 2021), https://taibbi.substack.com/p/meet-the-censored-hitler [https://perma.cc/9X7A-XVPV].

3. *Id.*

4. *Id.*

5. *See* Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INST. (Feb. 11, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels [https://perma.cc/JSM9-XPQL].

and harmful misinformation.[6] It is designed to produce an outcome with some practical consequences—such as remove or sustain, degrade, or otherwise reduce visibility—that will maintain Instagram's "frictionless, commercially successful product."[7]

Social media platforms may have a legitimate interest, and even a social duty, to address Holocaust denial and distortion and prevent the spread of disinformation.[8] Nonetheless, Marrota's post does not fit neatly under a rule against disinformation, and the system was most probably mistakenly triggered by the use of the name and/or the depiction of Hitler. Regardless of whether or not Instagram's AI system made the right call on this individual case, it failed to enable the important social dialogue between competing opinions regarding the legitimacy of Marrota's post and to give a voice to the different values at stake. Instead, like similar systems of speech moderation deployed by social media platforms, it is set to optimize removal of potentially harmful content, mathematically defined, while ignoring the "subtleties of different types of speech—differences between commentary and advocacy, criticism and incitement, [and] reporting and participation."[9]

Beyond the individual outcome, AI content moderation systems also exercise normative judgment, which is reflected in the way online speech norms are currently elaborated. While the terms of use of social media platforms often prohibit the spread of disinformation,[10] what is considered disinformation is embedded in the design of the system itself. This important normative judgment is opaque and therefore precludes any public engagement in checking and testing what these values are. However, if the legitimacy of Marrota's post was adjudicated in court, there would have been plenty of procedural room for deliberating its legitimacy and weighing its allegedly inciting or misleading potential against its historical-educational contribution. Different courts may have resolved the clash of values differently, and even if in the end, all adjudicators would have reached the same conclusion, the public could have still benefited from an open and transparent discussion about the values at stake and their normative

6. *See* Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41 (2020); *see also* KIRSTEN GOLLATZ ET AL., THE TURN TO ARTIFICIAL INTELLIGENCE IN GOVERNING COMMUNICATION ONLINE 3 (2018).

7. Taibbi, *supra* note 2.

8. A recent report by UNESCO found that nearly half of Holocaust-related content on Telegram either denied or distorted its history, while in moderated platforms, it was only 10-15%. *See* U.N. EDUC., SCI. & CULTURAL ORG., HISTORY UNDER ATTACK: HOLOCAUST DENIAL AND DISTORTION ON SOCIAL MEDIA 12, 27 (2022), https://unesdoc.unesco.org/ark:/48223/pf0000382159 [https://perma.cc/A9WZ-P6UJ].

9. *See* Taibbi, *supra* note 2.

10. Joan Donovan, *Here's How Social Media Can Combat the Coronavirus 'Infodemic'*, MIT TECH. REV. (Mar. 17, 2020), https://www.technologyreview.com/2020/03/17/905279/facebook-twitter-social-media-infodemic-misinformation/ [https://perma.cc/BY7A-THH6].

balance. Indeed, people could subsequently respond to the judicial resolution, support it, question its reasoning, or even press their representatives to change the law. In other words, the legal system would have facilitated ongoing public discussion, deliberation, and negotiation of speech norms pertaining to disinformation. As we further argue in this Article, although this process of democratic contestation is essential to democracy, its presence in algorithmic speech moderation is withering away.

A growing body of literature centers on the challenges raised by the deployment of automated tools to tackle potentially illegal or otherwise harmful content.[11] Many scholars have challenged the use of AI for speech governance on the ground of efficiency, questioning its ability to identify unwarranted content with precision and accuracy.[12] Others have questioned the legitimacy of using such systems by social media platforms.[13] The extraordinary power of digital platforms to shape online discourse and define the scope of freedom of expression has sparked a heated public debate over the concentration of speech governance power in the hands of a handful of private companies.[14]

Scholars have argued that the opaque, dynamic, and adaptive nature of AI tools creates significant barriers to public oversight[15] and

11. GOLLATZ ET AL., *supra* note 6; Henning Grosse Ruse-Khan, *Automated Copyright Enforcement Online: From Blocking to Monetization of User-Generated Content* 3-5 (PIJIP Rsch. Paper Series, Working Paper No. 51, 2020), https://digitalcommons.wcl.american.edu/research/51/ [https://perma.cc/3UDA-HWDZ]; Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated-Content*, NEW AM., https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content [https://perma.cc/K4H8-CRXM] (last updated July 22, 2019); Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech* 1, 3-4 (Nat'l Sec. Tech. & L. Working Grp., Aegis Series Paper No. 1902), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf [https://perma.cc/J5QT-XA33]; TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018); Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181 (2017).

12. *See* ALEXANDRE DE STREEL ET AL., ONLINE PLATFORMS' MODERATION OF ILLEGAL CONTENT ONLINE: LAWS, PRACTICES AND OPTIONS FOR REFORM 54 (2020).

13. *See* Amélie P. Heldt, *Upload-filters: Bypassing Classical Concepts of Censorship*, 10 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 56 (2019).

14. Jack M. Balkin, *Fixing Social Media's Grand Bargain* 1-2 (Nat'l Sec. Tech. & L. Working Grp., Aegis Series Paper No. 1814), https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf [https://perma.cc/7GQJ-89T3]; Thomas E. Kadri & Kate Klonick, *Facebook v.* Sullivan*: Public Figures and Newsworthiness in Online Speech*, 93 S. CAL. L. REV. 37, 39-40 (2019); KAREN KORNBLUH & ELLEN P. GOODMAN, SAFEGUARDING DIGITAL DEMOCRACY: DIGITAL INNOVATION AND DEMOCRACY INITIATIVE ROADMAP (2020); Moran Yemini, *Missing in "State Action": Toward a Pluralist Conception of the First Amendment*, 23 LEWIS & CLARK L. REV. 1149 (2020); Jonathan Zittrain, *How to Fix Twitter and Facebook,* ATLANTIC (June 9, 2022), https://www.theatlantic.com/ideas/archive/2022/06/elon-musk-twitter-takeover-mark-zuckerberg/661219/ [https://perma.cc/83TX-KWWZ].

15. FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 8 (2015); Maayan Perel & Niva Elkin-Koren,

threatens fundamental democratic principles.[16] Others have warned that AI tools could undermine autonomy and privacy as well as equality and accountability.[17] Scholars have also proposed policy measures to empower public oversight[18] and ensure compliance with civil rights in the use of AI-based speech governance.[19]

While many scholars have focused on individual and social harms generated by online content moderation, this Article focuses on the way AI systems generate speech norms, offering a new perspective on how speech governance by AI runs afoul of the democratic ideal of public participation and social deliberation.[20] It argues that speech governance by AI fails to sustain a normative space for contesting the limits of free speech, which is critical for democratic societies.

Contestation is central to the liberal democratic worldview.[21] Democratic contestation seeks to facilitate discursive interactions within civil society and to ensure that public debate enables citizens, as individuals and groups, to collectively form public opinion.[22] Three main elements underlie the notion of democratic contestation in the governance of speech: the first is the ability of individuals to object to speech norms and engage in ongoing critique about them, the second relates to public discourse being sufficiently open and inclusive to identify points of controversy over controversial speech norms, and the third is about facilitating a shared ground for voicing dissent and addressing it.[23]

Accountability in Algorithmic Copyright Enforcement, 19 STAN. TECH. L. REV. 473, 482 (2016).

16. J. Nathan Matias, Austin Hounsel & Melissa Hopkins, *We Tested Facebook's Ad Screeners and Some Were Too Strict*, ATLANTIC (Nov. 2, 2018), https://www.theatlantic.com/technology/archive/2018/11/do-big-social-media-platforms-have-effective-ad-policies/574609/ [https://perma.cc/4RBM-MTK4].

17. PASQUALE, *supra* note 15; Mireille Hildebrandt, *Saved by Design? The Case of Legal Protection by Design*, 11 NANOETHICS 307, 310 (2017).

18. *See, e.g.*, KORNBLUH & GOODMAN, *supra* note 14; Yifat Nahmias & Maayan Perel, *The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations*, 58 HARV. J. ON LEGIS. 145 (2021).

19. Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019); Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957 (2021).

20. *See infra* Part IV.

21. CLAUDE LEFORT, DEMOCRACY AND POLITICAL THEORY 231 (David Macey trans., 1988). As a practice of civil engagement, contestation is a critical component of democratic discourse. *See, e.g.*, WILLIAM SMITH, CIVIL DISOBEDIENCE AND DELIBERATIVE DEMOCRACY 9, 11 (2013).

22. *See infra* Section I.B.

23. Charles Girard, *Making Democratic Contestation Possible: Public Deliberation and Mass Media Regulation*, 36 POL'Y STUD. 283, 283 (2015). Charles Girard argues that contestable democracy should satisfy three conditions of contestability: it must be deliberative (creating a basis for contestation), it must be inclusive (creating a channel for the expression of dissenting voices), and it must be responsive (offering a forum where they could be met with a response).

In liberal democracies, the law facilitates democratic contestation by offering procedures and practices that are designed to hold together divided societies in the absence of deeper normative consensus. Fundamental democratic principles, such as separation of powers, judicial review, and the multiplicity of meanings generated through different interpretations by courts, provide the procedural framework that allows individuals and groups to pursue their diverse values in the democratic arena. The law further facilitates discursive interactions with extra-legal normative systems, such as custom and moral beliefs, while permitting diversity and inclusiveness.[24] Thus, the law upholds deliberative processes that create space for a normative dialogue over competing legitimate values, thereby sustaining legitimacy despite fundamental differences.

However, the shift to AI-based governance diminishes these basic democratic features. The use of ML to govern online speech rescinds any opportunity for civil negotiation over a multiplicity of meanings, as ML relies on probabilistic ex ante definitions and optimization dynamics.[25] The objective function of AI-based speech moderation is extracted automatically from the input data. Moreover, ML delivers data-driven speech norms without ensuring mechanisms that would enable ongoing deliberation over the tradeoffs they reflect. This (often efficient) mediation of disagreements over the legitimacy of speech by ML systems comes at the cost of withering important social space for democratic contestation over what constitutes legitimate speech and, more importantly, over how to decide the scope of legitimate discourse.

This Article argues that the design of AI-based systems of speech moderation should enable democratic contestation by making room for competing conceptions of tradeoffs and facilitating a common ground for negotiating positions, adjusting opinions, and making concessions.[26] Yet, attempting to ensure contestability by simply applying traditional legal procedures is doomed to be futile given the scope and scale of content moderation by AI.[27]

Therefore, to sustain democratic contestation in speech governance by AI, we propose a novel design intervention called *speech contestation by design*.[28] Inspired by the contestation mechanisms of the law, such as separation of powers and adversarial legal procedures, we suggest *separation of functions* and *contesting algorithms* as exemplary design features of AI systems of speech governance.

---

24. *See infra* Section II.C.

25. *See supra* note 20 and accompanying text; *infra* notes 311, 330 and accompanying text.

26. *See infra* notes 28, 330 and accompanying text.

27. Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526, 529 (2022).

28. *See infra* Section V.A.

This Article proceeds as follows. Part I provides the theoretical framework for our main argument. It highlights the central role of the public sphere in liberal democracies and explains how facilitating a democratic public discourse is the foundation of democratic societies. Further on, this Part elaborates on the notion of democratic contestation and its different inherent elements, contending that it is worryingly decreasing in the way our digital public sphere is currently governed. Part II turns to show how the law encourages democratic contestation. Specifically, the semantic (i.e., language based) nature of legal rules facilitates a multiplicity of meanings and flexibility in applying legal standards to different sets of circumstances, the distributed nature of law-making power enables diversity of meanings and multiple tradeoffs between free speech and conflicting values, and the way in which the evolution of legal norms is influenced by external normative systems further enables discursive negotiation over social norms. Then, Parts III and IV respectively explain how AI systems govern speech and why their current design fails to sustain sufficient space for democratic contestation. To fix this, Part V proposes to adopt speech contestation by design by embedding the democratic notions of separation of powers and adversarial legal procedures into the functional design features of AI systems of speech moderation. This Article concludes by proposing legal policy that may promote the integration of speech contestation by design in AI-based speech moderation systems.

I. THE PUBLIC SPHERE AND DEMOCRATIC CONTESTATION

A. The Public Sphere and Free Speech

The public sphere is a cornerstone of democracy.[29] It enables "the voicing of diverse views on any issue, the constitution of publicly-oriented citizens, the scrutiny of power and, ultimately, public sovereignty."[30] The public sphere is closely tied to democratic ideals that call for citizen participation in public affairs.[31] Such participation presumably enables a collective form of self-governance and, at the same time, also contributes to an individual's sense of existence and self-respect.[32]

29. *See generally* ALEXIS DE TOCQUEVILLE, DEMOCRACY IN AMERICA (Henry Reeve trans., 1990); Frederick Williams, *On Prospects for Citizens' Information Services*, *in* THE PEOPLE'S RIGHT TO KNOW: MEDIA, DEMOCRACY, AND THE INFORMATION HIGHWAY (Frederick Williams & John V. Pavlik eds., 1994).

30. Lincoln Dahlberg, *Rethinking the Fragmentation of the Cyberpublic: From Consensus to Contestation*, 9 NEW MEDIA & SOC'Y 827, 828 (2007).

31. Zizi Papacharissi, *The Virtual Sphere: The Internet as a Public Sphere*, 4 NEW MEDIA & SOC'Y 9, 10 (2002).

32. TOCQUEVILLE, *supra* note 29.

Inquiry and communication are viewed as the foundation of a democratic society, as they facilitate group deliberation over decisions made by a single authority.[33] The democratic ideal of self-governance by the people is thus grounded on the fundamental right of freedom of expression to ensure that citizens can share their ideas and thereby collectively form public opinion.[34]

As Robert Post argues, public discourse underpins all democratic theories.[35] Under a "participatory theory," wide participation in public discourse is a vehicle for enabling self-governance and constructing democratic legitimacy.[36] When citizens are free to engage in public discourse on matters of public concern, they are able to collectively contribute to the shaping of public policies and exercise their self-governance.[37] The participatory vision of democracy assumes access to information and the right of free deliberation by a well-informed citizenry. It presumes citizens have sufficient knowledge to independently form their opinion about public affairs and are capable of exercising their autonomy while collectively deciding their common destiny.[38] What counts as knowledge and relevance may also vary, and ensuring access to relevant knowledge requires not only reliable, but also diverse sources.[39] Therefore, wide participation in public discourse by all citizens not only seeks to safeguard the fundamental human right of free expression, but it is also instrumental to ensure that diverse views and opinions can be heard.

For liberal theorists, such as John Rawls, public debate should enable citizens to express their conceptions of the good and reach consensual solutions by reasoning based on shared principles.[40] Under participatory theories, wide participation in public discourse is a vehicle for enabling self-governance and constructing democratic legitimacy.[41] Critics of this approach are more skeptical of reasonable consensus as an ideal, raising concerns that it may stifle identity differences and conceal power relations. The purpose of democratic deliberation in an open society, they argue, is to allow these differences

33. *See generally* JOHN DEWEY, THE PUBLIC AND ITS PROBLEMS (1927).

34. Yemini, *supra* note 14, at 1192-93; Masses Publ'g Co. v. Patten, 244 F. 535, 540 (S.D.N.Y. 1917) ("[P]ublic opinion . . . is the final source of government in a democratic state.").

35. *See* Robert Post, *Reconciling Theory and Doctrine in First Amendment Jurisprudence*, 88 CALIF. L. REV. 2355 (2000).

36. Robert Post, *The Constitutional Status of Commercial Speech*, 48 UCLA L. REV. 1, 30 (2000).

37. Post, *supra* note 35, at 2367-68.

38. Ellen P. Goodman, *Digital Fidelity and Friction*, 21 NEV. L.J. 623, 625 (2021). *See generally* ERIC BARENDT, FREEDOM OF SPEECH (2d ed. 2007).

39. *See generally* YOCHAI BENKLER ET AL., NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS (2018).

40. *See generally* JOHN RAWLS, POLITICAL LIBERALISM (1993).

41. *See* Post, *supra* note 35, at 2371-72.

to be expressed and constantly renegotiated.[42] Nonetheless, whether under classic liberal theory, participatory theory, or the critique, participation in public discourse is essential for democracy.[43]

Participation in the public sphere could take different shapes and forms, ranging from directly voting on policy initiatives (referendum) to electing representatives that would promote a particular agenda through different governmental agencies.[44] Democratic theory assumes that citizens can take part in crafting social norms that apply to them not simply by going to the polls, but also by actively participating in the public sphere, namely deliberating on public affairs, and influencing the formation of norms. Contestation is a key feature in democratic participation, to which we turn next.

B. The Public Sphere and Democratic Contestation

Contestation over public policies plays a critical role in a democratic public sphere by providing legitimacy.[45] That is because, in reality, it is difficult to obtain the affirmative consent of all citizens to all public policies. The ability to contest may offer a second-best channel to proactive participation in public discourse. Contestability enables citizens to reflect their autonomous choice by objecting to policies with which they disagree.

What makes a democracy a form of self-ruling is often not the ability to manifest choice regarding each policy which may affect our lives, but is rather the ability to contest decisions and possibly revise them. Therefore, contestation, as the ability of citizens to oppose a particular decision, is viewed as essential for legitimacy.[46]

Another function of contestation is to restrain power by creating channels for challenging power.[47] Contestation as an institutional design principle, for instance, aims to restrain the domination of coercive power held by the government to safeguard civil liberties by dispersing power in competing institutions.[48] Such institutional design is reflected by the democratic principle of separation of powers, whereby government responsibilities are divided between competing branches of government, each overseeing the other.

---

42. *See generally* CHANTAL MOUFFE, THE DEMOCRATIC PARADOX (2000).

43. Maria Ferretti & Enzo Rossi, *Pluralism, Slippery Slopes and Democratic Public Discourse*, 60 THEORIA 29, 29 (2013).

44. Post, *supra* note 35, at 2367-68.

45. *See, e.g.*, CHARLES TILLY & SIDNEY TARROW, CONTENTIOUS POLITICS 8 (2d ed. 2007).

46. *See* Girard, *supra* note 23 ("[P]ublic policies are legitimate not simply because of their substantive content or procedural origin, but because they can be contested, and sometimes revised, even after they have been enacted.").

47. Seymour Martin Lipset, *The Indispensability of Political Parties*, 11 J. DEMOCRACY 48, 48 (2000).

48. Benjamin A.T. Graham et al., *Safeguarding Democracy: Powersharing and Democratic Survival*, 111 AM. POL. SCI. REV. 686 (2017).

Contestation could take different forms.[49] Sometimes it refers to the ability to dispute or object to a decision or an action taken by an authority,[50] such as in adversarial legal disputes, arbitrations, or appeal procedures. Contestation may also take the form of a social or political act to challenge a position or an ideology, where individuals and groups could discursively express disapproval of norms which govern society. Indeed, political contestation can be seen as "vital to reinvigorating what is left of the anarchic political energies of the public sphere and pushing or 'encouraging' institutions to pay more attention to the points of view and demands articulated by the great variety of more or less organized actors in the public sphere."[51] Social protests which have turned into social movements, such as Me Too or climate change, are exemplary.[52] Political contestation might also take the form of a legal intervention, such as petitioning against the Texas Abortion Act.[53] Arguably, a common feature in all of these acts of contestation is disagreement regarding the desirability of some norms which govern our society, some different perceptions regarding the meaning of such norms, and often the need to make choices between competing values and meanings in a legitimate manner.

Contestation as a practice of civil engagement is a critical component of democratic discourse.[54] Democratic contestation, on which we focus in this paper, seeks to facilitate discursive interactions within civil society to ensure that public debate enables citizens, as individuals and groups, to collectively form public opinion. This understanding of democratic contestation entails several elements: it is discursive; it must be open and inclusive of diverse voices; and it must be deliberative, offering a shared ground for collectively deciding conflicting views. We further explain these elements below.

First, democratic contestation is discursive. As Antje Wiener puts it: "[T]he concept's analytical utility lies in understanding the distinct meanings of contestation as both a social practice of merely objecting to norms (principles, rules, or values) by rejecting them or refusing

49. Antje Wiener, *A Theory of Contestation—A Concise Summary of Its Argument and Concepts*, 49 POLITY 109, 109 (2017).

50. Marco Almada, *Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems*, *in* PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2019).

51. Robin Celikates, *Digital Publics, Digital Contestation: A New Structural Transformation of the Public Sphere?*, *in* TRANSFORMATIONS OF DEMOCRACY: CRISIS, PROTEST AND LEGITIMATION (Robin Celikates et al. eds., 2015).

52. *See* ME TOO., https://metoomvmt.org/ [https://perma.cc/ED7V-3JXE] (last visited Sept. 23, 2023); *Solutions for the Planet*, CLIMATE FOUND., https://www.climatefoundation.org/ [https://perma.cc/2MPX-CYBV] (last visited Sept. 23, 2023).

53. Brianna Coates, *Fight Against Texas' New Abortion Law*, CHANGE.ORG (May 19, 2021), https://www.change.org/p/governor-greg-abbott-fight-against-texas-new-abortion-law [https://perma.cc/XH77-2QP7].

54. *See* SMITH, *supra* note 21.

to implement them[] and as a mode of critique through critical engagement in a discourse about them."[55] Second, democratic contestation aims to facilitate a diversity of voices. Contestation as a practice of democratic civil engagement[56] should enable all citizens to question and express their objection to political decisions, fundamental norms, or ideologies. Openness and inclusiveness of public discourse aim to ensure that it is capable of facilitating better social choices by disclosing flaws, underlying points of controversy, and helping focus public debates on the social choices to be made.[57]

Third, democratic contestation presumes a shared ground for voicing dissent and addressing it.[58] Democratic contestation further seeks to promote reasoning.[59] Arguably, citizens who cannot learn to be critical, or to reason on matters of public affairs, are also likely to be less autonomous. Engaging in a social dialogue may enable members of society to shape their own opinions. The deliberation of social norms may therefore involve an explicit articulation of the norm and the underlying values it invokes.[60] Deliberation might further enable individuals to tweak their opinions to find common grounds with those of others, persuade one another, or otherwise switch opinions when confronted with persuasive arguments. Thus, democratic contestation must offer some space for socially negotiating different views and collectively deciding priorities and tradeoffs.

Indeed, the diversity of opinions (the second element) and the need to establish a shared grounding of public opinion (the third element) might seem contradictory. Arguably, the right and ability to contest may introduce more opinions and perspectives to the public debate, thus including more voices in public debate and promoting diversity. Yet, simply voicing diversified opinions might be insufficient for serving the functions of the public sphere. Without paying attention to the *way* public discourse is structured, simply enabling more opinions could only enhance partisanship, sectarianism, and polarization in society.[61]

To assist members of society to collectively form public opinion, public discourse must not only be dialogic (allow persuasion) and enable deliberation, but must also facilitate the formation of some

55. *See* Wiener, *supra* note 49.

56. *See* SMITH, *supra* note 21.

57. Michael Coppedge, Angel Alvarez & Claudia Maldonado, *Two Persistent Dimensions of Democracy: Contestation and Inclusiveness*, 70 J. POL. 632 (2008).

58. *See* Girard, *supra* note 23. Charles Girard argues that contestable democracy should satisfy three conditions of contestability: it must be deliberative (creating a basis for contestation), it must be inclusive (creating a channel for the expression of dissenting voices), and it must be responsive (offering a forum where they could be met with a response).

59. *Id.*

60. SMITH, *supra* note 21; Iris Marion Young, *Activist Challenges to Deliberative Democracy*, 29 POL. THEORY 670, 685-88 (2001).

61. As discussed below in Section I.C.

shared grounding. A common ground for deliberation is necessary to highlight which issues are rendered public, what values should be weighed, and where social negotiation of norms should take place. Democratic contestation could help create such common ground by highlighting competing framings, underscoring different points of departure in public debate, and disclosing the points of controversy.

Importantly, a common ground does not entail agreement on substantive norms but strives towards a shared framing of issues and procedures for addressing conflicts. It further involves mechanisms for resolving conflicts between competing values, which are based on acknowledging the standing of political opponents and the legitimacy of their (often conflicting) ideas.[62] Democratic discourse would seek to facilitate multiple principles and diversity in resolving conflicts of values by allowing for the coexistence of different *kinds* and different *conceptions* of values. Such pluralism enables liberal democracies to thrive and to develop morally and societally: keeping society whole while maintaining nuances and differences.

All in all, an important feature of democratic contestation in the public sphere is to facilitate discursive interactions within civil society and to ensure that public debate enables citizens, as individuals and groups, to collectively form public opinion. Democratic contestation should therefore provide a framework for participation, deliberation, and reasoning to facilitate a dialogic public discourse.

C. The Digital Public Sphere

Democratic institutions and legal procedures aim at facilitating democratic contestation, as further demonstrated in Part II. The rise of a digital public sphere introduces, however, new types of challenges to the democratic contestation ideal, to which we turn next.

At the beginning of the century, the emergence of the Internet has been seen as introducing a more egalitarian public sphere, offering citizens new opportunities to encounter and directly engage with a wide diversity of positions.[63] The digital public sphere as a newly decentralized network has raised high hopes that it would promote democratic discourse, where users could freely share their expressions with billions of other users around the world.[64] By exploiting a variety of online communication mechanisms, different actors could articulate

62. *See* SMITH, *supra* note 21; *see also* Young, *supra* note 60.

63. Dahlberg, *supra* note 30, at 828.

64. MANUEL CASTELLS, COMMUNICATION POWER 87-88 (2009); CLAY SHIRKY, HERE COMES EVERYBODY: THE POWER OF ORGANIZING WITHOUT ORGANIZATIONS 171 (2008); YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 176-77 (2006); Elettra Bietti, *A Genealogy of Digital Platform Regulation*, 7 GEO. L. TECH. REV. 1, 1, 12 (2023).

and critique the validity of different claims.[65] Yet, in fact, the digital public sphere shows a worrying departure from the democratic ideal.[66]

Public discourse in modern times resides on digital platforms.[67] A handful of social media platforms, like Facebook, YouTube, and Twitter, have become digital public squares where opinions, ideas, and preferences are shaped.[68] They dominate the online conversation, undermining the mitigating power of competitive pressures.[69] These digital platforms that operate in multisided markets[70] deploy various digital tools on users to harvest data and extract revenues from selling users' profiles for targeted advertising or other data-driven products and services.[71] Some of these tools may have a divisive influence on public discourse.[72] The viral spread of extremist content, reinforced by algorithmic "filter bubbles" and online "echo-chambers," have all contributed to deepening social divides.[73]

---

65. Dahlberg, *supra* note 30, at 828.

66. *See, e.g.*, Lincoln Dahlberg, *The Habermasian Public Sphere: Taking Difference Seriously?*, 34 THEORY & SOC'Y 111 (2005); Lincoln Dahlberg, *The Internet and Discursive Exclusion: From Deliberative to Agonistic Public Sphere Theory*, *in* RADICAL DEMOCRACY AND THE INTERNET: INTEGRATING THEORY AND PRACTICE (Lincoln Dahlberg & Eugenia Siapera eds., 2007); Graham Murdock & Peter Golding, *Dismantling the Digital Divide: Rethinking the Dynamics of Participation and Exclusion*, *in* TOWARD A POLITICAL ECONOMY OF CULTURE: CAPITALISM AND COMMUNICATION IN THE TWENTY-FIRST CENTURY (Andrew Calabrese & Colin Sparks eds., 2004).

67. Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018); Amélie P. Heldt, *Merging the Social and the Public: How Social Media Platforms Could Be a New Public Forum*, 46 MITCHELL HAMLINE L. REV. 997 (2020).

68. Moran Yemini, *The New Irony of Free Speech*, 20 COLUM. SCI. & TECH. L. REV. 119, 122, 125 (2018); Kadri & Klonick, *supra* note 14.

69. *See generally* Elettra Bietti, *Consent as a Free Pass: Platform Power and the Limits of the Informational Turn*, 40 PACE L. REV. 310, 311 (2019) (analyzing how notice and consent aspects of media platform's ToS provide inadequate protection to the average user).

70. DAVID S. EVANS & RICHARD SCHMALENSEE, MATCHMAKERS: THE NEW ECONOMICS OF MULTISIDED PLATFORMS 98 (2016).

71. Jean-Charles Rochet & Jean Tirole, *Two-Sided Markets: A Progress Report*, 37 RAND J. ECON. 645, 650 (2006); Max Freedman, *How Businesses Are Collecting Data (and What They're Doing with It)*, BUS. NEWS DAILY, https://www.businessnewsdaily.com/10625-businesses-collecting-data.html [https://perma.cc/4UTS-MD6C] (last updated May 30, 2023).

72. Axel Bruns, *Filter Bubble*, INTERNET POL'Y REV., Nov. 29, 2019, at 1; Richard Fletcher & Rasmus Kleis Nielsen, *Are People Incidentally Exposed to News on Social Media? A Comparative Analysis*, 20 NEW MEDIA & SOC'Y 2450 (2018); Nicolas M. Anspach, *The New Personal Influence: How Our Facebook Friends Influence the News We Read*, 34 POL. COMMC'N 590 (2017).

73. *See* Bruns, *supra* note 72. The Mozilla Foundation, for instance, has recently investigated the negative ways in which YouTube's recommendation algorithm impacted the wellbeing of YouTube's users. This investigation revealed that YouTube's algorithm is recommending videos that violate their own terms of use and harm people. Mozilla's report notes that YouTube's recommendation system plays an "outsized part" in radicalization as it steers users towards radical content, and "once people are 'in' the rabbit hole," the recommendation algorithm offers them "more extreme ideas." MOZILLA FOUND., YOUTUBE REGRETS: A CROWDSOURCED INVESTIGATION INTO YOUTUBE'S RECOMMENDATION ALGORITHM 5 (2021).

Especially, the central role of democratic contestation as enabling deliberation over competing views is withering away. Scholars question "whether the myriad of diverse views that exist online are actually intersecting, and thus the extent to which online interactions actually involve any significant problematization and contestation of positions and practices."[74] In reality, online discourse is fragmented. Digital conversation involves like-minded individuals with shared identity, leading to what Sunstein names "enclaves for communication."[75] Instead of enabling users to confront opposing views, the Internet has become "a breeding ground for polarization" and "extremism."[76] As Sunstein explains, following deliberation with others of shared identity, "people are likely to move toward a more extreme point in the direction to which the group's members were originally inclined."[77] Polarization could therefore lead to hostility and even violence, which threatens our democratic public sphere.[78]

As we further explain in Part III, the infrastructure of the digital public sphere, and particularly how it is governed, diminishes democratic contestation while sustaining this polarization. Speech governance by AI fails to offer a rescue to democratic contestation. As we further show in Part IV, the process of shaping the norms that govern online speech is currently driven solely by data.

D. Sustaining Democratic Contestation in Times of Social Divides

The transition to speech governance by AI is taking place at a moment of crisis in liberal democracies, where societies are deeply divided over the practical meaning of freedom of expression and its legitimate boundaries.[79] Following two decades of flourishing freedom of expression, boosted by the Internet, free speech in recent years is under siege. Data collected by Freedom House shows that free speech has been declining both in authoritarian regimes and in liberal democracies.[80] Public debate reflects disagreements on many issues of substance, such as whether governments should intervene in markets, how to balance national security and human rights, or what measures should be taken to ensure public health during a global pandemic.

---

74. Dahlberg, *supra* note 30, at 828.

75. Anupam Chander, *Whose Republic*, 69 U. CHI. L. REV. 1479, 1489 n.45 (2002) (reviewing CASS R. SUNSTEIN, REPUBLIC.COM (2001)).

76. SUNSTEIN, *supra* note 75, at 71.

77. *Id.* at 65.

78. Dahlberg, *supra* note 30, at 830.

79. *See, e.g.*, Samuel Earle, *The 'Culture Wars' Are a Symptom, Not the Cause, of Britain's Malaise,* GUARDIAN (May 31, 2021, 3:00 PM), https://www.theguardian.com/commentisfree/2021/may/31/culture-wars-symptom-not-cause-britains-malaise [https://perma.cc/WZ3F-CFK5].

80. *See* FREEDOM HOUSE, FREEDOM IN THE WORLD (2019).

Yet, in recent years, it seems that disputes are no longer confined to the substance of speech but have now extended also to its legitimacy, namely, whether particular expressions should be allowed at all.

There is a growing disagreement regarding the boundaries of legitimate speech that are worthy of protection against undue restraints. "No-platforming" and boycotts on college campuses, designed to prevent particular speakers from being heard, are viewed by some as censorship and by others as legitimate protest.[81] Angry tweets are framed by some as abusive attempts to silence legitimate speech and by others as a reasonable attempt to hold speakers accountable.[82] What some see as selective enforcement by social media platforms intended to silence conservative speakers[83] is perceived by others as an inadequate response to a viral spread of toxic expressions and dangerous incitements to violence.[84]

Indeed, liberal democracies have become deeply divided over the value of freedom of expression and the boundaries of legitimate speech and how conflicts over those boundaries should be resolved.[85] While

81. *See* Mary Anne Franks, *The Miseducation of Free Speech*, 105 VA. L. REV. ONLINE 218, 220 (2019).

82. Nesrine Malik, *The Myth of the Free Speech Crisis: How Overblown Fears of Censorship Have Normalized Hate Speech and Silenced Minorities*, GUARDIAN (Sept. 3, 2019, 1:00 PM), https://www.theguardian.com/world/2019/sep/03/the-myth-of-the-free-speech-crisis [https://perma.cc/7B9H-4BWX].

83. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (June 2, 2020). *See generally* Evelyn Douek, *Trump Is a Problem That Twitter Cannot Fix*, ATLANTIC (May 27, 2020), https://www.theatlantic.com/ideas/archive/2020/05/twitter-cant-change-who-the-president-is/612133/ [https://perma.cc/R3LE-7CF9].

84. Douek, *supra* note 83 (emphasizing the argument that adding a fact-check link to former president Trump's tweets was insufficient, and that instead, the offending tweets must come down).

85. Consider, for instance, the public controversy sparked by the so-called Harper's letter illustrates the deep disagreement in liberal societies over the value of free expression and its legitimate boundaries. In a "Letter on Justice and Open Debate" published in Harper's Magazine, more than 150 prominent artists, academics, and journalists applauded recent calls for social justice but at the same time voiced concern over illiberal voices, an intolerant climate, a stifling atmosphere, and "cancel culture." The signatories warned against the "moral attitudes and political commitments that tend to weaken our norms of open debate and toleration of differences in favor of ideological conformity." The letter voiced concern that "[t]he free exchange of information and ideas, the lifeblood of a liberal society, is daily becoming more constricted." The signatories called for preserving space for good faith disagreement and argued that "[t]he way to defeat bad ideas is by exposure, argument, and persuasion, not by trying to silence or wish them away." *A Letter on Justice and Open Debate,* HARPER'S MAG. (July 7, 2020), https://harpers.org/a-letter-on-justice-and-open-debate/ [https://perma.cc/TZ5N-L45G]. In an open counter letter, critics claimed that free speech has served mostly the powerful, arguing that "[t]he signatories, many of them white, wealthy, and endowed with massive platforms," speak from a position of privilege and refuse to accept the reality of a diversifying industry—"one that's starting to challenge institutional norms that have protected bigotry." The signatories of the counter letter argued that "[u]nder the guise of free speech and free exchange of ideas, the letter appears to be asking for unrestricted freedom to espouse their points of view free from consequence or criticism." *A More Specific Letter on Justice and Open Debate*, OBJECTIVE (July 10, 2020), https://objectivejournalism.org/2020/07/a-more-specific-letter-on-justice-and-open-debate/ [https://perma.cc/BF7S-ASQ2].

most agree that threats of violence fall outside the constitutional protection of freedom of expression, there is wide disagreement as to what counts as violent speech. Some believe that words which are offensive to certain disadvantaged groups are in themselves inherently violent and should be banned.[86] At the same time, what counts as offensive has increasingly become subjective—it is offensive if it offends me.[87] These developments reflect a profound departure from the notion of free speech as *free from any restraint*, which was once the norm in the United States.[88] There is no longer a consensus over the meaning of the right to free speech, as reflected in the Voltairean phrase "I disapprove of what you say, but I will defend to the death your right to say it."[89]

Hence, especially today, liberal democracies must preserve a democratic space for deliberating the disagreements in society regarding the scope of free speech. It is necessary to sustain a procedural framework for engaging in a social dialogue over the development of speech norms that shape our digital public sphere. Next, we turn to show how speech governance by law pursues this goal.

II. DEMOCRATIC CONTESTATION IN SPEECH GOVERNANCE BY LAW

In liberal democracies, the law seeks to facilitate a space for democratic contestation by offering procedures and practices which are designed to hold together divided societies in the absence of deeper normative consensus. As Rawls frames it, in procedural terms, democratic pluralism endorses the idea of agreeing on "the political procedures of democratic government."[90] For instance, fundamental democratic principles, such as separation of powers, judicial review, and the multiplicity of meanings generated through different interpretations by courts, provide the procedural framework that allows individuals and groups to pursue their diverse values in the democratic arena. This *thin* liberal approach, sometimes called "liberalism of fear," seeks to foster "peaceful coexistence among competing and incommensurable ways of life" and even incommensurable moral commitments.[91]

86. *Hate Speech and Hate Crime*, AM. LIBR. ASS'N (Dec. 12, 2017), https://www.ala.org/advocacy/intfreedom/hate [https://perma.cc/RT7G-TNH3].

87. Margaret Martin, Censorship in the Age of Identity Politics (2020) (unpublished manuscript) (available at https://pos.direito.ufmg.br/wp-content/uploads/2020/09/Paper-Margaret-Martin-Censorship-in-the-Age-of-Identity-Politics.pdf [https://perma.cc/93ND-6RQM]). *See generally* JOEL SIMON, THE NEW CENSORSHIP: INSIDE THE GLOBAL BATTLE FOR MEDIA FREEDOM (2019).

88. Cass R. Sunstein, *Free Speech Now*, 59 U. CHI. L. REV. 255, 258-60 (1992).

89. S.G. TALLENTYRE, THE FRIENDS OF VOLTAIRE 199 (1906).

90. RAWLS, *supra* note 40, at 159.

91. Nathan Oman, *Contract Law and the Liberalism of Fear*, 20 THEORETICAL INQUIRIES L. 381, 402 (2019).

One obvious way by which the legal system facilitates democratic contestation is by dispersing the power to decide and interpret norms among competing institutions.[92] The lawmaking power is vested in different branches of the government, and each branch can generate legal norms, sometimes with contradictory implications.[93] Democratic contestation is further advanced by the way the law interconnects with extra-legal normative systems while making room for diversity and inclusiveness. While the law, according to its internal logic, is exclusive, it may be challenged by different sources of normative principles that govern human behavior,[94] such as informal understandings which are embedded in culture.[95] While these sets of norms coexist, intertwine, and sometimes conflict, they do not necessarily displace the conception of the law as a unified and coherent system. Rather, this perspective considers state law as but one form of law within a context of normative multiplicity.[96]

Finally, the law itself involves mechanisms that facilitate a plurality of meanings and at the same time sustain a common ground of contestation.[97] As explained by Reichman:

> [T]he official norms and procedures governing the conduct of state agencies[] do not form a monolithic singular, coherent entity which we may call "the law;" rather, the different substantive norms, procedures, and institutions empowered to settle factual and normative disputes in a state form a collage of multiple facets of "law[,]" some of which are in tension with each other.[98]

Different rules of conflict inform the adjudicator which facet of the law should apply to reach a legal resolution in any given case.[99] These rules may arrange the different facets of the law in a hierarchical order, limit the application of each facet to a certain domain of the legal universe, or do both.[100]

Below, we expand on these features, which facilitate democratic contestation in the context of speech governance by law. In Part IV we will later demonstrate how these features are lacking in the current design of AI-based governance of speech. This analysis will set the

92. Gordon R. Woodman, *Ideological Combat and Social Observation: Recent Debate About Legal Pluralism*, 42 J. LEGAL PLURALISM & UNOFFICIAL L. 21, 37, 46 (1998).

93. *See* Amnon Reichman, Neo-Formalism as Formal Legal Pluralism (2022) (unpublished manuscript) (on file with author).

94. *Id.*

95. LAUREN B. EDELMAN & MARC GALANTER, LAW: THE SOCIO-LEGAL PERSPECTIVE 604 (James D. Wright ed., 2d ed. 2015).

96. Margaret Davies, *The Ethos of Pluralism*, 27 SYDNEY L. REV. 87, 96 (2005).

97. *Id.* ("Positive law can be regarded as inherently, irreducibly plural—full of gaps, contradictions, unresolved histories, counter-narratives and, most pertinently, composed of multiple dimensions and layers.").

98. Reichman, *supra* note 93, at 13.

99. *Id.* at 51.

100. *Id.*

ground for our proposal to incorporate contestation by design, which is inspired by the rule of law, into the systems of AI-based speech moderation.

A. One Norm, Multiple Interpretations

An important feature of law, which facilitates ongoing social negotiation of speech norms, is its semantic and distributed nature.[101] The use of language to shape behavior enables legal norms to encompass different, often conflicting, meanings, ascribed simultaneously by different legal agents. This, in turn, creates a critical space for negotiating values and adjusting the meaning of norms over time and space. This is especially the case concerning legal principles.[102] Consider, for instance, the legal definition of copyright infringement. The Copyright Act of 1976 provides that any reproduction of a protected work of authorship is copyright infringement.[103] But what if the alleged infringer reproduces only some portion of a protected work? It is unclear how much of the original work must be reproduced to establish infringement. Therefore, courts have developed the "substantial similarity" test to determine infringement.[104] According to the Second Circuit, "[t]his test judges whether, in the eyes of the ordinary observer, there is a substantial similarity between the protected work and the allegedly infringing work."[105] Other courts use different tests.[106] Most importantly, the law enables different meanings of "substantial similarity" to coexist and allows ad hoc determinations of infringement to be made down the road.

Norms are sometimes intentionally kept broad and ambiguous by lawmakers, allowing them to sustain different meanings, in order to bridge diverse interests and goals. Legal standards make use of open-ended terms, such as "reasonable," "fair," or "due diligence," which facilitate "sophisticated methods of social control."[107] Unlike rules, which explicitly define legal consequences that result from easily ascertainable facts, open-ended standards allow the judge to define

101. Jerzy Wroblewski, *Semantic Basis of the Theory of Legal Interpretation*, 6 LOGIQUE ET ANALYSE 397, 397 (1963).

102. *See infra* notes 107-10 and accompanying text.

103. 17 U.S.C. § 501(a) (1982).

104. Amy B. Cohen, *Masking Copyright Decisionmaking: The Meaninglessness of Substantial Similarity*, 20 U.C. DAVIS L. REV. 719, 733-34 (1987).

105. *Id.* at 722.

106. Jeannette Rene Busek, *Copyright Infringement: A Proposal for a New Standard for Substantial Similarity Based on the Degree of Possible Expressive Variation*, 45 UCLA L. REV. 1777, 1778-79 (1998).

107. George C. Christie, *Vagueness and Legal Language*, 48 MINN. L. REV. 885, 889 (1964).

preconditions for the legal consequences when applying the norm.[108] Indeed, "rules precede the incident (ex ante), while when setting standards the judge formulates the norm upon its application, namely, after the incident has taken place (ex post)."[109] Vague general standards can evolve over time through a series of particular applications and change in content as the nature of society changes.[110]

The "fair use" standard in U.S. copyright law is a classic example. Under fair use, one who makes unauthorized use of a protected work in a fair manner does not infringe the exclusive rights of the copyright owner.[111] As noted, acknowledging the "endless variety of situations and combinations of circumstances that can rise" and wanting to avoid "freez[ing] the doctrine in the statute, especially during a period of rapid technological change,"[112] Congress adopted a notoriously vague fair use provision.[113] The statutory provision of fair use provides a nonexclusive list of possibly fair purposes of use,[114] along with a list of four factors derived from case law that must be taken into account to determine fair use.[115] Based on this vague language, judges must carve out exceptions for otherwise infringing uses after weighing a set of factors on a case-by-case basis. Hence, the judge must not only determine whether certain preconditions exist in the case at hand, but must also exercise judicial discretion to define which factors are relevant for determining that the fair use doctrine applies. As a result, the doctrine may "be applied to a variety of uses" and in different contexts, "including to uses and in contexts that Congress may not

---

108. Kathleen M. Sullivan, *The Supreme Court, 1991 Term—Foreword: The Justices of Rules and Standards*, 106 HARV. L. REV. 22, 121 (1992).

109. Niva Elkin-Koren & Orit Fischman-Afori, *Rulifying Fair Use*, 59 ARIZ. L. REV. 161, 168 (2017).

110. Christie, *supra* note 107, at 890.

111. 17 U.S.C. § 107 (2006).

112. H.R. REP. No. 94-1476, at 66 (1976), *as reprinted in* 1976 U.S.C.C.A.N. 5678, 5680.

113. Jason Mazzone, *Administering Fair Use*, 51 WM. & MARY L. REV. 395, 400 (2009).

114. 17 U.S.C. § 107 specifies that "the fair use of a copyrighted work, including such use by reproduction in copies . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright."

115. Section 107 provides:

> In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.

Id.

have anticipated at the time it passed the law."[116] Such open-ended standards not only facilitate flexibility and dynamism in applying legal norms to specific cases, but also serve as a modus vivendi,[117] allowing social agreement on high-level principles, removed from immediate conflicting interests, while deferring disagreements to be resolved down the road. In other words, vagueness facilitates the ongoing deliberation of meanings, which "allows man to exercise general control over his social development without committing himself in advance to any specific concrete course of action."[118]

Moreover, the nature of legal norms is neither inherent nor intrinsic. Instead, the attributes of rules and standards are subject to interpretation by courts.[119] Different theories of legal interpretation—such as textualism, legislative intentionalism, and purposivism—all seek to discover the meaning of law.[120] Judges often soften rules and insert more discretionary judgment at the moment of application by introducing exceptions or applying broad interpretations that extend beyond the literal meaning of the rule.[121] In that sense, even rules that ought presumably to be strict in their application are subject to judicial interpretation. Consider, for instance, the exclusive right of reproduction accorded to the owner of a copyrighted work.[122] Although the law provides that a copyright owner has the exclusive right "to reproduce the copyrighted work in copies or phonorecords,"[123] courts must determine which actions are considered reproductions for the purpose of copyright liability. Through judicial interpretation, courts can adjust the meaning of the rule to meet changing circumstances. The Second Circuit, for instance, interpreted "reproduction" by imposing two requirements: first, the copied work must be embodied in a medium, and second, it must remain embodied "for a period of more than transitory duration."[124] Accordingly, the court concluded that reproduction in an online buffer for a brief period of 1.2 seconds did not meet the duration requirement, and the statutory meaning of "reproduction" therefore did not apply.[125]

116. Mazzone, *supra* note 113, at 400-01.

117. DAVID MCCABE, MODUS VIVENDI LIBERALISM: THEORY AND PRACTICE 133 (2010).

118. Christie, *supra* note 107, at 890.

119. Reichman, *supra* note 93, at 52.

120. RICHARD H. FALLON ET AL., HART AND WECHSLER'S THE FEDERAL COURTS AND THE FEDERAL SYSTEM 652-56 (Robert C. Clark et al. eds., 7th ed. 2015).

121. FREDERICK F. SCHAUER, PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LAW AND IN LIFE 170 (1991).

122. 17 U.S.C. § 106(1) (2012).

123. *Id.*

124. Cartoon Network LP v. CSC Holdings, Inc., 536 F.3d 121, 126-27, 130 (2d Cir. 2008).

125. *Id.* at 130.

B. Contestation as an Institutional Design Principle

An important institutional design principle in law, which facilitates contestation, relates to the distributed power of lawmaking. Specifically, the democratic principle of "separation of powers" allocates irreducible lawmaking power to the three branches of government—the legislative branch, the executive branch, and the judicial branch.[126] In common law constitutional democracies, this creates "three legal regimes, each organized around a set of constitutive elements that govern, as a matter of ideal types, the engagement with the regime."[127] As argued by Huq and Michaels, this democratic principle, which is often associated with "Madisonian resistance to tyranny (as reflected in the separation of powers) and the corresponding commitment to pluralism (as reflected in the diversification of powers)[,] should be reconceived to reflect not just concern about literal, corporeal tyranny, but also about the tyranny of a single norm."[128]

In practice, diverse separation of powers values are contested and ultimately realized in a multitude of venues.[129] As noted, the "three branches [of government] serve as devices through which a larger, pluralistic normative vision can be channeled and, ultimately, vindicated."[130] Separation of powers is thus "intended to simultaneously advance and harmonize diverse and conflicting normative ends."[131]

Consider, as an example, the legal debate over hate speech. On the one side are those who "understand hate speech to be a means of perpetuating systematic discrimination and oppression of minority groups."[132] They perceive " 'freedom of speech' as a screen that protects racism, homophobia, misogyny, and other forms of discrimination," urging that "the equality values of the Fourteenth Amendment must not be sacrificed in the name of the First Amendment."[133] On the other side are those who claim that "defining a category of 'hate speech' will be difficult" and that

126. *See, e.g.*, Michael L. Yoder, Note, *Separation of Powers: No Longer Simply Hanging in the Balance*, 79 GEO. L.J. 173, 173 (1990).

127. *Id.*

128. Aziz Z. Huq & John D. Michaels, *The Cycles of Separation-of-Powers Jurisprudence*, 126 YALE L.J. 346, 381 (2016) (emphasis omitted).

129. *Id.*

130. *Id.* at 382.

131. *Id.*

132. Charlotte H. Taylor, *Hate Speech and Government Speech*, 12 U. PA. J. CONST. L. 1115, 1117 (2010).

133. *Id.*

"allowing the government to suppress a particular viewpoint, even one that is unequivocally condemned by a majority of the population, opens the door for further government censorship."[134]

The federal government is bound by a constitutional commitment to free speech: the First Amendment Free Speech Clause provides that Congress shall make "no law . . . abridging the freedom of speech."[135] However, the executive branch may sometimes give preference to the opposite position as exemplified by the ordinance enacted by St. Paul, Minnesota against hate speech.[136] This ordinance created a distinct, separate criminal misdemeanor for symbolic conduct of a hatred nature, prohibiting the display of a symbol which one knows or has reason to know "arouses anger, alarm or resentment in others on the basis of race, color, creed, religion or gender."[137] In *R.A.V. v. City of St. Paul*, R.A.V. was charged under this ordinance for burning a cross in the middle of the night on a black family's front lawn.[138] The Minnesota Supreme Court found the ordinance applicable and constitutional while narrowing its scope to cover only unprotected "fighting words."[139] Unprotected speech was viewed as regulable speech not fully protected by the First Amendment.[140] Later, however, the ordinance was struck down by the Supreme Court as unconstitutional.[141] The Court viewed differently the meaning of "unprotected," finding that while unprotected speech such as fighting words could be regulated because of its "constitutionally proscribable content," the government cannot "regulate them based on hostility, or favoritism, towards a nonproscribable message they contain."[142]

Two decades later, in the year 2021, state legislators have introduced more than 100 bills aiming to regulate how social media companies handle users' content.[143] Two of those have become actual laws in Florida and Texas—Republican states fighting against the alleged censorship of conservative viewpoints—that sought to prohibit

134. *Id.*

135. U.S. CONST. amend. I.

136. ST. PAUL, MINN., LEG. CODE § 292.02 (1990).

137. *Id.* ("Whoever places on public or private property a symbol, object, appellation, characterization or graffiti, including, but not limited to, a burning cross or Nazi swastika, which one knows or has reasonable grounds to know arouses anger, alarm or resentment in others on the basis of race, color, creed, religion or gender commits disorderly conduct and shall be guilty of a misdemeanor.").

138. R.A.V. v. City of St. Paul, 505 U.S. 377, 379-80 (1992).

139. *See In re* R.A.V., 464 N.W.2d 507, 509-10 (Minn. 1991), *rev'd sub nom.* R.A.V. v. City of St. Paul, 505 U.S. 377 (1992).

140. *Id.* at 509-11.

141. *R.A.V.*, 505 U.S. at 377.

142. *Id.* (emphasis omitted).

143. Rebecca Kern, *Push to Rein in Social Media Sweeps the States*, POLITICO (July 1, 2021, 4:30 AM), https://www.politico.com/news/2022/07/01/social-media-sweeps-the-states-00043229 [https://perma.cc/3S83-5T7A].

tech platforms from ousting political candidates.[144] The United States Court of Appeals for the Eleventh Circuit ruled that the Florida law restricting social media was largely unconstitutional.[145] As to the Texas law, the Supreme Court blocked it,[146] though the United States Court of Appeals for the Fifth Circuit had previously unblocked the law,[147] and it still faces a lawsuit from two tech industry groups.[148] On the other side of the political map, New York, a Democratic state, has enacted a new law requiring social media networks to make it possible for individuals to report hate speech on the platforms in a publicly accessible way; failure to comply with the law may expose platforms to a fine of $1000 a day.[149]

The legal conversation between federal law, state legislatures, and state courts shapes the boundaries of free speech in a discursive fashion that reflects different conceptions of free speech as articulated by different political viewpoints. Since each branch of the government may generate speech norms, and such norms may be in conflict, disputes as to which norm governs in each case are unavoidable.[150] These disputes facilitate deliberation over different meanings of the law. Often, these different meanings coexist. Indeed, "in a federal state, we can conceive of norms diverging along geographically-organized state structures. Such divergence can be conceived of as plurality: the norms governing the same activities are different in different places within the same country."[151] Thus, the law is not monolithic.

C. A Common Ground for Negotiating Diverse Meanings

Speech governance by legal norms is discursive, making room for multiple meanings not only internally, but also externally. The meaning of norms might be informed by other kinds of normative systems, such as custom, culture, and religion, which pluralize its function.[152] Indeed, while formal norms are produced by legislators and interpreted by administrative agents and courts, the law is a social practice that cannot be understood outside a social context.[153]

144. Fla. SB 7072 (2021); Tex. HB 20 (2021).

145. NetChoice, LLC v. Att'y Gen., 34 F.4th 1196, 1231 (11th Cir. 2022).

146. NetChoice, LLC v. Paxton, 142 S. Ct. 1715 (2022).

147. NetChoice, LLC v. Paxton, No. 21-51178, 2022 WL 1537249 (5th Cir. May 11, 2022).

148. Taylor Hatmaker, *Supreme Court Pauses Controversial Texas Social Media Law*, TECHCRUNCH (May 31, 2022, 6:46 PM), https://techcrunch.com/2022/05/31/texas-social-media-law-supreme-court-hb20/ [https://perma.cc/B8LH-7X8B].

149. N.Y. GEN. BUS. LAW § 394-ccc (McKinney 2022).

150. Reichman, *supra* note 93, at 12.

151. *Id.*

152. Sally Falk Moore, *Law and Social Change: The Semi-Autonomous Social Field as an Appropriate Subject of Study*, 7 LAW & SOC'Y REV. 719, 721 (1973).

153. Malcolm M. Feeley, *The Concept of Laws in Social Science: A Critique and Notes on an Expanded View*, 10 LAW & SOC'Y REV. 497, 501 (1976).

As Ronald Beiner explains, judgment is impossible unless there are "underlying grounds of judgment which human beings, qua members of a judging community, share, and which serve to unite in communication even those who disagree (and who may disagree radically)."[154]

The theoretical ideal of legislators who act in the public interest, administrative agencies that enforce clear-cut rules, and judges who apply legal norms in a technically impartial manner[155] has been challenged by numerous law and society scholars.[156] Rather, legislators, judges, administrative agencies, and lawyers all adjust and interpret the law in light of their social context.[157] Put differently, " 'the spirit of law' . . . is not simply invented at the top but is transformed, challenged, and reinvented in local practices that produce a plural legal culture."[158]

The U.S. obscenity jurisprudence is exemplary. The governing standard for obscenity is based on three criteria:

> (a) whether "the average person, applying contemporary community standards" would find that the work, taken as a whole, appeals to the prurient interest; (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.[159]

"Appeal[s] to the prurient interest" and "patent offensiveness," however, are both to be judged with reference to contemporary community standards.[160] For that reason, "[n]o definition of obscenity could ever be formulated with sufficient clarity that it would target only constitutionally unprotected speech."[161] Put differently, until the Supreme Court, applying vague standards, finds a specific material to be obscene, no one can ever say with certainty

154. RONALD BEINER, POLITICAL JUDGMENT 142 (1983) (emphasis omitted).

155. *See* William Baude & Stephen E. Sachs, *The Law of Interpretation*, 130 HARV. L. REV. 1079 (2017).

156. EDELMAN & GALANTER, *supra* note 95, at 606.

157. *See* Stephen M. Feldman, *Supreme Court Alchemy: Turning Law and Politics into Mayonnaise*, 12 GEO. J.L. & PUB. POL'Y 57 (2014).

158. Barbara Yngvesson, *Inventing Law in Local Settings: Rethinking Popular Legal Culture*, 98 YALE L.J. 1689, 1693 (1989).

159. Miller v. California, 413 U.S. 15, 24 (1973) (quoting Kois v. Wisconsin, 408 U.S. 229, 230 (1972)).

160. *See, e.g.*, Ashcroft v. ACLU, 535 U.S. 564, 576 n.7 (2002); *see also* Pope v. Illinois, 481 U.S. 497, 500 (1987).

161. Bret Boyce, *Obscenity and Community Standards*, 33 YALE J. INT'L L. 299, 319 (2008).

that it is so.[162] Therefore, "obscenity" and "pornography" should be treated as "placeholders for contested meaning" that should always be regarded "as if there were quotation marks around them."[163]

Applying a legal norm to a particular set of circumstances also involves the exercise of discretion and thereby incorporates other normative systems, such as ethics and culture, in merging legal rules and non-legal principles.[164] Legal principles are therefore fluid and dynamic, facilitating continuous change in response to ongoing negotiation of meaning and validity by social actors who are themselves subject to entwining normative systems.[165] This interpretative nature of legal norms leaves further room for diversity of meanings at all levels.

At the same time, however, legal procedures, institutions, and rights offer a common ground for negotiating these diverse meanings and even contesting their framing.[166] The law evolves on the basis of particularity, depending on the proficiency and level of the court deciding the case, the surrounding circumstances, the characteristics of the specific clash being resolved, and the characteristics of the authorized decisionmaker exercising interpretive power. While seeking coherence, the law makes room for a broad spectrum of tradeoffs between competing values to coexist. It does so by delegating interpretative power to a distributed network of judges acting within a distributed system of courts.[167] By enabling agreement on high-level principles while leaving room for ongoing social negotiation and interpretation of legal norms, this system allows for resolution of clashes between fundamental rights and basic values on a case-by-case basis. Different resolutions of a clash between two similar values or interests could be and are in fact possible, thus sustaining the capacity of the law to evolve and adjust its normative structure.

---

162. *Id.*

163. Amy Adler, *What's Left?: Hate Speech, Pornography, and the Problem for Artistic Expression*, 84 CALIF. L. REV. 1499, 1506, 1508 (1996). Boyce likewise argues that the meanings of obscenity and pornography are contested and "historically contingent." Boyce, *supra* note 161, at 304.

164. Robert Post, *Theorizing Disagreement: Reconceiving the Relationship Between Law and Politics*, 98 CALIF. L. REV. 1319, 1343 (2010) (arguing that unlike politics, which presumes disagreement between members of the polity, "[l]aw is a social practice that presumes agreement" and must therefore enable ways "to tame, channel and resolve ongoing, persistent disagreement").

165. Yngvesson, *supra* note 158. Yngvesson contends that viewing legal culture in a "dynamic way can . . . explain popular consciousness as a force contributing to the production of legal order rather than as simply an anomaly or a pocket of consciousness 'outside' of law, irrelevant to its maintenance and transformation." *Id.* at 1693.

166. BEINER, *supra* note 154, at 143 ("Judgment implies a community that supplies common grounds or criteria by which one attempts to decide.").

167. *See, e.g.*, Felix Frankfurter, *Distribution of Judicial Power Between United States and State Courts*, 13 CORNELL L. REV. 499 (1928).

In sum, speech governance by law allows citizens to contest meanings on a shared ground, towards collectively deciding conflicting views, while often disagreeing.

The transition to AI in speech governance by social media platforms undermines some of these fundamental features of governance by legal norms, as explained next.

III. GOVERNING SPEECH BY AI

Frictionless flows of information have become a signature trait of the digital economy.[168] Information flows face no national borders, no mismatches between technical standards, no physical boundaries, and very low transaction costs—all of which have made the sharing of content and personal data smooth and swift. Any content posted by a user on social media could potentially become available to millions of other users worldwide. This type of viral distribution has undoubtedly generated economic efficiency and promoted important social values, giving rise to social movements such as *#MeToo*.[169] At the same time, some content shared by users may be harmful. This has posed new challenges to digital platforms that host such content, forcing them to undertake different strategies of content moderation.[170] Below we describe the essence of speech moderation by AI.

A. The Rise of Speech Moderation by AI

Social media platforms create a space where content originated by users can be shared, thereby enabling individuals and groups to connect around content generated by users.[171] Content moderation is thus the core function provided to users of social media.[172] Content moderation refers to practices such as classifying content posted by users by determining whether such content can or should be published, with whom it can be shared, and under what conditions. As observed by Grimmelman, content moderation is "the governance mechanism[] that structure[s] participation in a community to facilitate cooperation and prevent abuse."[173]

---

168. *See* Goodman, *supra* note 38.

169. Stephanie Nicholson et al., *A Platform for Empowerment: Social Media and the Social Diffusion of the #MeToo Movement*, *in* NEW PERSPECTIVES ON CRITICAL MARKETING AND CONSUMER SOCIETY 199, 206-07 (Elaine L. Ritch & Julie McColl eds., 2021).

170. Robert Gorwa et al., *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y, Jan.-June 2020, at 1.

171. Giovanni De Gregorio, *Democratising Online Content Moderation: A Constitutional Framework*, 36 COMPUT. L. & SEC. REV. 1, 1-2 (2020); Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1353 (2017).

172. GILLESPIE, *supra* note 11, at 21.

173. James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 47 (2015) (emphasis omitted).

Platforms engage in two types of content moderation. The first is intended to match content with users' interests and preferences ("content curation"). The second is intended to ensure compliance with community standards and legal duties ("content filtration").[174] The core business of platforms is to match content with users and facilitate users' engagement with content (i.e., viewing, reacting, and responding to content) for the purpose of lengthening the amount of time users spend on the platform, which in turn increases the platforms' advertising income.[175] The greater the traffic on the platform, as measured in the number of new users and the time spent on the platform by existing users, the more revenues are generated for the platform.[176]

The matching of content with viewers is made possible by algorithms, which are used to predict users' preferences based on their previous behavior and that of similar others and to direct content toward users who are most likely to view and potentially respond to it. For instance, when deciding which movies to recommend to subscribers, Netflix may compare data collected on that subscriber's viewing history with the profiles of millions of others as a means to predict the individual's viewing preferences or how likely they are to try new content.[177] Similarly, Facebook curates users' news feeds,[178] and YouTube sets its recommendation system in accordance with users' predicted preferences.[179] These efforts are all designed to enhance online engagement and thus to catalyze further traffic on the platform and maximize advertising.[180]

The second type of content moderation aims at tackling potentially harmful content uploaded by users. These practices include "the

174. Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV. 857, 875, 880 (2020); Amélie P. Heldt, *Content Moderation by Social Media Platforms: The Importance of Judicial Review*, *in* CONSTITUTIONALISING SOCIAL MEDIA 251 (Edoardo Celeste, Amélie P. Heldt & Clara Iglesias Keller eds., 2022).

175. *See* Mathew Ingram, *How Google and Facebook Have Taken over the Digital Ad Industry*, FORTUNE (Jan. 4, 2017, 1:30 PM), https://fortune.com/2017/01/04/google-facebook-ad-industry/ [https://perma.cc/V3XW-FB9X].

176. *See* Sarah T. Roberts, *Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation*, 23 FIRST MONDAY 3 (2018) (explaining that users' content can be considered "the currency by which users are engaged as consumers and producers on social media sites").

177. *See How Netflix's Recommendation System Works,* NETFLIX, https://help.netflix.com/en/node/100639 [https://perma.cc/6P3V-BZ62] (last visited Sept. 23, 2023).

178. *See generally* GILLESPIE, *supra* note 11.

179. Cristos Goodrow, *On YouTube's Recommendation System*, YOUTUBE OFF. BLOG (Sept. 15, 2021), https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/ [https://perma.cc/94SM-CLNG].

180. *See* Karen Hao, *YouTube Is Experimenting with Ways to Make Its Algorithm Even More Addictive*, MIT TECH. REV. (Sept. 27, 2019), https://www.technologyreview.com/2019/09/27/132829/youtube-algorithm-gets-more-addictive/ [https://perma.cc/J6VE-QNEA].

screening, evaluation, categorization, approval[,] or removal/hiding of online content according to relevant communications and publishing policies . . . to support and enforce positive communications behavior online[] and to minimize aggression and anti-social behavior."[181] In this context, content moderation strategies seek to ensure that content complies with appropriate norms, either internal (i.e., community guidelines) or external (i.e., regulatory restraints), by filtering, blocking, downgrading, or removing inappropriate content.[182] The rise of visibility sanctions, such as delisting and downranking, whereby content is not entirely removed but rather its visibility to users is reduced,[183] is blurring the distinction between content curation and content filtration.

In the past, platforms relied on human moderators to screen content uploaded to social media.[184] With the amount of content growing exponentially, platforms were forced to supplement and even replace human review with automated systems.[185] The massive scale of content hosted by platforms, and the speed at which content must be assessed and dealt with, pose an enormous logistic challenge and by themselves may be sufficient to make the case for shifting to AI in speech moderation. Automated flagging and removal are of gigantic scale. For instance, during Q1 2022, more than 90% (3,544,195) of the 3,882,684 videos removed by YouTube for violating its Community Guidelines were flagged by automated systems.[186]

Moreover, platforms which offer livestreaming services must swiftly classify and remove any harmful time-sensitive livestreamed content, such as terrorist attacks, murders, or sexual assaults.[187] On top of this, platforms face business and political challenges that push them to deploy AI in content moderation.[188] Human content moderation practices have attracted criticism over the political bias of

181. *See* Terry Flew et al., *Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance*, 10 J. DIGIT. MEDIA & POL'Y 33, 40 (2019).

182. *See* Martin Husovec, *The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown? Which Is Superior? And Why?*, 42 COLUM. J.L. & ARTS 53, 59 (2018).

183. *See generally* Kelley Cotter, *Playing the Visibility Game: How Digital Influencers and Algorithms Negotiate Influence on Instagram*, 21 NEW MEDIA & SOC'Y 895 (2019).

184. *See generally* Hector Postigo, *Emerging Sources of Labor on the Internet: The Case of America Online Volunteers*, 48 INT'L REV. SOC. HIST. 205, 205 (2003) (mentioning among the duties of the American Online Volunteers (AOV) were to enforce the Terms of Use agreement).

185. *See* Gorwa et al., *supra* note 170; *see also* CAMBRIDGE CONSULTANTS, USE OF AI IN ONLINE CONTENT MODERATION 16 (2019).

186. *See YouTube Community Guidelines Enforcement*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/youtube-policy/removals?hl=en [https://perma.cc/5JXW-RKEB] (last visited Sept. 23, 2023).

187. *See* Gorwa et al., *supra* note 170.

188. *See generally* Tarleton Gillespie, *Content Moderation, AI, and the Question of Scale*, BIG DATA & SOC'Y, July-Dec. 2020, at 1-4.

human reviewers[189] and led to a public outcry over the distressing work conditions of content moderators, which were argued to be harmful to their mental health.[190] Facebook recently agreed to pay $52 million to settle a class action brought by human moderators, who claimed that they experienced post-traumatic stress disorder from reviewing content on Facebook's sites.[191]

The alleged political bias of human content moderators has sparked vivid political debate. Arguably, algorithmic editorial processes might be more neutral compared to the human beings who traditionally determined the content people encountered, namely the editors of newspapers and television news programs.[192] Indeed, notwithstanding some concerns regarding intentional or subliminal bias in the programming of algorithms,[193] algorithms are often conceived as more neutral and objective than humans.[194]

The adoption of automated measures is also a result of increasing regulatory pressures, including the expanding liability of online platforms for potentially harmful content posted by their users.[195] Recent legislative and regulatory provisions in Europe now encourage platforms to act promptly against the dissemination of unlawful content.[196] In a similar vein, the United States 2021 Appropriations Act directed the Federal Trade Commission (FTC) to provide recommendations on the use of AI against specified online harms,

---

189. See, e.g., the Executive Order signed by President Trump, entitled "Preventing Online Censorship," in which he claimed that online platforms function as "a 21st century equivalent of the public square," accused them of engaging in "selective censorship" that harms public discourse, and instructed federal agencies to take action to protect against such alleged censorship. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020).

190. *See generally* SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA 209 (2019).

191. *See* Bobby Allyn, *In Settlement, Facebook to Pay $52 Million to Content Moderators with PTSD*, NPR (May 12, 2020, 10:52 PM), https://www.npr.org/2020/05/12/854998616/in-settlement-facebook-to-pay-52-million-to-content-moderators-withptsd [https://perma.cc/7L6A-MNJB].

192. *See* Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1025 (2017).

193. *See* PASQUALE, *supra* note 15; Jonathan Zittrain, *Facebook Could Decide an Election Without Anyone Ever Finding Out*, NEW REPUBLIC (June 1, 2014), https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering [https://perma.cc/MGG2-TWCS].

194. *See generally* Nizan G. Packin, *Consumer Finance and AI: The Death of Second Opinions?*, 22 N.Y.U J. LEGIS. & PUB. POL'Y 319 (2020); Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035 (2018); Bloch-Wehba, *supra* note 6.

195. *See generally* Niva Elkin-Koren, Yifat Nahmias & Maayan Perel, *Is It Time to Abolish Safe Harbor? When Rhetoric Clouds Policy Goals*, 31 STAN. L. & POL'Y REV. 1 (2020).

196. *See, e.g.*, Article 17(1), European Parliament (March 26, 2019); *see also* Deutscher Bundesrat: Drucksachen [BR-Drs.] 536/17 (Ger.); Directive 2017/541, of the European Parliament and of the Council of 15 March 2017 on Combating Terrorism and Replacing Council Framework Decision 2002/475/JHA and Amending Council Decision 2005/671/JHA, 2017 O.J. (L 88) 6, 9.

including fraud, deepfakes, harassment, hate crimes, terrorist content, and election-related disinformation.[197]

The COVID-19 pandemic has further accelerated the transition from human moderators to automated measures. Allowing content moderators to work remotely involved new privacy and security challenges.[198] The pandemic has forced major social media platforms, including Facebook,[199] YouTube,[200] and Twitter,[201] to reduce their use of human reviewers and rely primarily on automated systems.[202] However, a recent report prepared by the Congressional Research Service maintains that the growing reliance of social media platforms on automated content moderating systems during the COVID-19 pandemic led to an increase in errors, including both the mistaken removal of legitimate content and failures to remove illicit content.[203]

All in all, many platforms today deploy AI systems both to optimize the matching of users' content and to improve the speedy detection of potentially harmful content, to filter unwarranted content before it is posted, to identify and track similar content, and to block access to it or remove it from the platform. As we have argued elsewhere,[204] the different functions performed by digital platforms in content moderation—curating personalized content for targeted advertising and filtering allegedly illicit content—are all embedded in the same system. As we further explain below, AI-driven content moderation performs its functions through the labeling of users and content, application programming interfaces (API), learning patterns, and

197. *See* Consolidated Appropriations Act, 2021, H.R. 133, 116th Cong. § 1501 (2020) (enacted).

198. *See* Elizabeth Dwoskin & Nitasha Tiku, *Facebook Sent Home Thousands of Human Moderators Due to Coronavirus. Now the Algorithms Are in Charge*, WASH. POST (Mar. 24, 2020, 5:55 PM), https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/ [https://perma.cc/EGX5-GVJV]; *see also* Shannon Bond, *Facebook, YouTube Warn of More Mistakes as Machines Replace Moderators*, NPR (Mar. 31, 2020, 5:06 AM), https://www.npr.org/2020/03/31/820174744/facebook-youtube-warn-of-more-mistakes-as-machines-replacemoderators [https://perma.cc/TLJ6-P92J].

199. *See* Kang-Xing Jin, *Keeping People Safe and Informed About the Coronavirus*, META (Dec. 18, 2020), https://about.fb.com/news/2020/12/coronavirus/#keeping-our-teams-safe [https://perma.cc/83HV-H2HL].

200. *Protecting Our Extended Workforce and the Community*, YOUTUBE OFF. BLOG (Mar. 16, 2020), https://blog.youtube/news-and-events/protecting-our-extended-workforce-and/ [https://perma.cc/8MXE-8H88].

201. Vijaya Gadde & Matt Derella, *An Update on Our Continuity Strategy During COVID-19*, TWITTER BLOG (Apr. 1, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html [https://perma.cc/UGV3-43QZ].

202. *See* Jack Goldsmith & Andrew Keane Woods, *Internet Speech Will Never Go Back to Normal*, ATLANTIC (Apr. 25, 2020), https://www.theatlantic.com/ideas/archive/2020/04/what-covid-revealed-about-internet/610549/ [https://perma.cc/W9LN-RJDD].

203. JASON A. GALLO & CLARE Y. CHO, CONG. RSCH. SERV., SOCIAL MEDIA: MISINFORMATION AND CONTENT MODERATION ISSUES FOR CONGRESS 7 (2021).

204. Elkin-Koren & Perel, *supra* note 174, at 857, 884.

software. Consequently, decisions on removal of speech, for (public) law enforcement purposes, are driven by the same data, algorithms, and optimization logic which also underlie all other functions performed by digital platforms.

Next, we shift focus to some of these features to further understand how they shape the decisionmaking process pertaining to the scope of permissible content.

B. Speech Governance by AI

AI systems make use of algorithms and data to identify patterns and make predictions. There is no consensus over the definition of AI, and the term is commonly used to describe a broad array of techniques.[205] Currently, many content moderation systems make use of ML techniques, which enable systems to "learn" how to perform a certain task by training on vast volumes of data.

"[M]achine learning," as described by David Lehr and Paul Ohm, "refers to an automated process of discovering correlations (sometimes alternatively referred to as relationships or patterns) between variables in a dataset, often to make predictions or estimates of some outcome."[206] The algorithm is set to optimize an objective function (namely, the mathematical expression of the algorithm's goal).[207] For instance, the objective function of a system designed to predict copyright infringement might be to correctly classify infringing content (namely, uploaded content that is substantially similar to the copyrighted content). Optimizing this goal means maximizing accurate predictions, or, alternatively, minimizing inaccurate ones—the percentage of uploaded works incorrectly identified as infringing (false positives) or non-infringing (false negatives). Eventually, such systems attain the capacity to analyze new data and make predictions by drawing on their prior learnings.[208]

ML systems installed in the upload filters of social media are deployed to detect illicit speech, such as hate speech, terrorist propaganda, and copyright infringements.[209] For instance, Scribd, a subscription-based digital library of e-books and audiobooks, employs a system called BookID to generate a digital fingerprint for each book based on semantic data (e.g., word counts, letter frequency, and phrase

---

205. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 404 (2017) ("There is no straightforward, consensus definition of artificial intelligence."); Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORNELL L. REV. 287, 293-94 (2020) ("[T]here is something exceptional about robots and AI that make them exceptionally difficult to define.").

206. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

207. *Id.*

208. *Id.* at 672.

209. GOLLATZ ET AL., *supra* note 6, at 3.

comparisons).[210] Texts uploaded to Scribd are scanned by BookID, and content which matches any BookID fingerprint is blocked.[211] Similarly, Amazon's Project Zero uses ML to continuously scan product listing updates and to proactively remove suspected counterfeits, based on logos, trademarks, and key data provided by its partnering brands.[212] Another instance of intellectual property enforcement via ML is YouTube's Content ID. Using a digital identifying code, Content ID can detect and notify right holders whenever a newly uploaded video matches a work that they own. Right holders can then choose to block or remove the content, share information, or monetize the content.[213]

All systems that rely on automated data-driven decisionmaking processes rely on datafication—namely, a choice embedded in the system as to which data to collect and to record.[214] As aptly argued by Nissenbaum, data is not simply a raw resource "lying about awaiting collection"; rather, it is "constructed or created from the signals of countless technical devices and systems."[215] Typically, AI-based content moderation systems have four main features. First, they have a system for *labeling* data as either legitimate or unwarranted. Second, they work through a *predictive model*, which predicts whether any given content is illicit based on features learned in the training model. Third, they use automated decisionmaking to choose and undertake the *action* to be performed (e.g., post, recommend, remove, block, or filter). Finally, a key feature of ML content moderation systems is a recursive *feedback loop*. Once trained, these systems enter an organic process of continual learning. Content identified as illicit is fed back into the model so that it will be detected the next time the system runs.[216]

---

210. *About the BookID™ Copyright Protection System*, SCRIBD HELP CTR., https://support.scribd.com/hc/en-us/articles/360037497152-About-the-BookID-Copyright-Protection-System [https://perma.cc/W2Q2-AMQ5] (last visited Sept. 23, 2023).

211. *Id.*

212. Dharmesh M. Mehta, *Amazon Project Zero*, AMAZON (Feb. 28, 2019), https://blog.aboutamazon.com/company-news/amazon-project-zero [https://perma.cc/5MFG-ZDEJ].

213. Katharine Trendacosta, *Unfiltered: How YouTube's Content ID Discourages Fair Use and Dictates What We See Online*, EFF (Dec. 10, 2020), https://www.eff.org/wp/unfiltered-how-youtubes-content-id-discourages-fair-use-and-dictates-what-we-see-online [https://perma.cc/2GRA-8RHC]; Matthew Sag, *Internet Safe Harbors and the Transformation of Copyright Law*, 93 NOTRE DAME L. REV. 499, 541-42 (2017); *see also* Perel & Elkin-Koren, *supra* note 15, at 510.

214. *See* Katherine J. Strandburg, *Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 5, 5-6 (Julia Lane, Victoria Stodden, Stefan Bender & Helen Nissenbaum eds., 2016).

215. Helen Nissenbaum, *Must Privacy Give Way to Use Regulation?*, *in* DIGITAL MEDIA AND DEMOCRATIC FUTURES 255, 264-65 (Michael X. Delli Carpini ed., 2019) (emphasis omitted).

216. Gorwa et al., *supra* note 170, at 4-6.

Content moderation can be based on either *supervised learning* or *unsupervised learning*.[217] Supervised learning involves training the algorithm with previously labeled data designed to classify different types of content.[218] Labeling refers to the recording, aggregating, tagging, and coding of data into a format that can be used for training and data analytics. This can be done internally by the platform that operates the content moderation system, or it can be outsourced.[219] The system may be given a large set of (probably) correct answers to the system's task (labeled content), and it learns to answer new cases in a similar way.[220] Hence, a system meant to detect hate speech might be trained through a set of posts where content amounting to hate speech was distinguished from the rest of the content. Likewise, to train the system to weed out terrorist propaganda, training data might include images labeled "Islamic State propaganda" and the like alongside images labeled "legitimate."[221] With sufficient training data, the system should learn to distinguish terrorist propaganda from everything else. Systems using digital hash technology may also learn to identify content that is similar to the labeled content.[222] Digital hash technology converts images or videos into a hash ("digital signature"), which is a significantly smaller file than the original and thus a more convenient file to analyze.[223] Some hashing techniques (especially "perceptual hashing") may be resistant to alterations, thereby enabling the identification of not exact matches, such as resized images or images with minor color alterations.[224] This enables the screening of online content, ex post or ex ante, against a database of predefined illicit content.[225] For instance, a system could be trained to identify images showing the use of firearms or to identify matches in files sharing similar metadata. Every new piece of content that is identified updates the database and becomes embedded in future screenings of the system.

---

217. Lehr & Ohm, *supra* note 206, at 673 ("Supervised algorithms are given a labeled outcome variable (alternatively called an output or response variable) representing the true values to be predicted on the basis of input data.").

218. Gorwa et al., *supra* note 170, at 5.

219. For instance, in an effort to address misinformation, in December of 2016 Facebook launched its Third-Party Fact-Checking Program, whereby independent fact-checking partner organizations examine content on the site. Content defined as misinformation is labeled as such, and users' ability to share it may be restricted. *Meta's Third-Party Fact-Checking Program*, META, https://www.facebook.com/journalismproject/programs/third-party-fact-checking [https://perma.cc/34WP-DD3M] (last visited Sept. 23, 2023).

220. CAMBRIDGE CONSULTANTS, *supra* note 185, at 19.

221. However, the system is not confronted with borderline material that was allowed due to specific circumstances but would have been banned were the circumstances different.

222. *See generally* EVAN ENGSTROM & NICK FEAMSTER, THE LIMITS OF FILTERING: A LOOK AT THE FUNCTIONALITY & SHORTCOMINGS OF CONTENT DETECTION TOOLS (2017).

223. CAMBRIDGE CONSULTANTS, *supra* note 185, at 12.

224. Gorwa et al., *supra* note 170, at 4.

225. *Id.* at 4-5.

Unsupervised learning, by contrast, does not make predictions based on pre-labeled content but instead seeks to cluster content based on certain shared characteristics. Unsupervised Domain Adaptation for Hate Speech Detection, for instance, identifies hate speech sentences where the hate speech terms can be distinguished from their surrounding sentence context to create a template for domain adaptation.[226] The algorithm then identifies the template in generic sentences to slot in hate speech and convert it into hate speech in a new domain.[227] To create a domain-adapted corpus, a sequential tagger is trained on the labeled data in the source domain so that the tagger is able to identify hate speech content terms and surrounding sentence context templates.[228] Later, the tagger is applied "to unlabeled data in the target domain to derive a lexicon of hate terms in the target domain."[229]

AI systems deployed in content moderation could focus on classifying the content alone, checking whether it matches certain classifiers that render it likely to be unwarranted. Systems might also attain the capacity to analyze personal data, drawing on the poster's prior behavior in order to make predictions regarding the risk potentially posed by content based on the identity of the poster or the content creator. For instance, a study by the Center for Countering Digital Hate (CCDH) published in March 2021 showed that the majority of COVID-19 anti-vaccine misinformation and conspiracy theories posted on Facebook, Instagram, and Twitter earlier that year originated from just twelve people.[230] Similarly, a German-based conspiracy group was found to coordinate a loose network of conspiracy-laced groups that helped to drive a series of anti-lockdown protests across Australia, which then turned into violent clashes.[231] ML tools could be applied to track the spread of disinformation and identify its sources.[232]

At the same time, processing personal data related to users might also enable ML systems to differentiate between different contexts

226. Sheikh Muhammad Sarwar & Vanessa Murdock, *Unsupervised Domain Adaptation for Hate Speech Detection Using a Data Augmentation Approach*, *in* PROCEEDINGS OF THE 16TH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA 852, 853 (2022).

227. *Id.*

228. *Id.*

229. *Id.*

230. *See* CTR. COUNTERING DIGIT. HATE, THE DISINFORMATION DOZEN: WHY PLATFORMS MUST ACT ON TWELVE LEADING ONLINE ANTI-VAXXERS (2021).

231. Christopher Knaus & Michael McGowan, *Who's Behind Australia's Anti-Lockdown Protests? The German Conspiracy Group Driving Marches*, GUARDIAN (July 26, 2021), https://www.theguardian.com/australia-news/2021/jul/27/who-behind-australia-anti-covid-lockdown-protest-march-rallies-sydney-melbourne-far-right-and-german-conspiracy-groups-driving-protests [https://perma.cc/S83D-U3NZ].

232. *Corneliu* Bjola & Ilan Manor, *Combating Online Hate Speech and Anti-Semitism* (Oxford Digit. Dipl. Rsch. Grp., Working Paper No. 4, 2020).

that may also affect the legitimacy of use. For instance, some use of copyright materials by students or teachers for the purpose of learning might be considered fair use.[233]

C. Speech Norms by AI

Automated speech moderation by AI does not only affect the rights of each individual user to post content on digital platforms.[234] Since social media platforms constitute a digital public square, limits on speech also affect the rights of others to learn from that speech. Speech moderation is a form of governance that generates norms, shapes practices, and coordinates the behavior of social actors.[235]

How do algorithms govern speech? Speech regulation, in a broad sense, defines the scope of permissible speech through social and legal norms.[236] AI introduces a new type of governance, which is based on dynamic and adaptive decisionmaking processes driven by data, correlations, and predictions.

The scope of permissible speech on digital platforms is typically defined in legal terms, which are listed in the platforms' Terms of Service (ToS). These contractual provisions often incorporate more detailed guidelines (e.g., Facebook's Community Standards or YouTube's Community Guidelines).[237] Users who accept a platform's ToS enter a contract whereby they are required to adhere to these norms when using the platform to share content.[238] In practice, however, it is ML systems that define the scope of permissible and unlawful speech. These definitions are later embedded in upload filters of social media that are often set to enforce these norms, thus effectively providing an operational definition of permissible use through the technical details.[239]

AI systems govern speech by creating speech affordances—that is, determining which content remains available and which content is

233. Niva Elkin-Koren, *Fair Use by Design*, 64 UCLA L. REV. 1082 (2017).

234. Yifat Nahmias & Maayan Perel, *The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations*, 58 HARV. J. ON LEGIS. 145, 149 (2021).

235. This broad view of governance is not limited to command-and-control by state agencies, but rather covers a whole range of regulatory interventions by various social actors. *See generally* David Levi-Faur, *Regulation and Regulatory Governance*, *in* HANDBOOK ON THE POLITICS OF REGULATION 3 (David Levi-Faur ed., 2011).

236. *See* Klonick, *supra* note 67, at 1603.

237. *Facebook Community Standards*, META, https://transparency.fb.com/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F [https://perma.cc/RRB5-7GZ8] (last visited Sept. 23, 2023); *Community Guidelines*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ [https://perma.cc/2GW2-LG4T] (last visited Sept. 23, 2023).

238. Edoardo Celeste, *Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?*, 33 INT'L REV. L. COMPUT. & TECH. 122, 123 (2019).

239. *See* Maayan Perel, *Digital Remedies*, 35 BERKELEY TECH. L.J. 1, 26-27 (2020).

removed[240]—and how content might be shared (e.g, "like" or "retweet").[241] More precisely, through their technical definitions of particular features and their respective weights, ML systems effectively define whether a certain piece of content —be it image, text, or video—is classified as illegitimate speech that is subject to removal. ML systems can also shape the spread of speech, determining which users can view it and how often.[242] YouTube's restricted mode, for instance, is an optional setting that tags potentially mature or objectionable content and prevents users with restrictions enabled from viewing it.[243] Some algorithms can also limit who can participate in online conversations (e.g., by requiring verification of the user's online identity or by suspending accounts).[244]

This is also the case where systems are set to detect illegal content based on law, such as child pornography, inciting materials, counterfeit products,[245] or copyright infringements. Determinations of judicial and semi-judicial issues regarding such illegal content depend on the technical implementation of ML content moderation systems.[246] For instance, the threshold of substantial similarity in copyright law or the particular score that defines a piece of content as obscenity must be embedded in the ML system, which then makes a purely mechanical judgment.

In this context, note that speech governance via AI does not merely apply existing norms, thereby simply reflecting existing values and tradeoffs. In discerning between content that is permissible and content that is banned, automated content moderation systems also craft norms and shape users' behavior. Consider, for instance, a proactive tool recently announced by YouTube.[247] The new tool, called "Checks," is based on YouTube Content ID. It allows users to screen videos they intend to upload before actually doing so, to check whether these videos contain copyrighted material and whether they comply

240. *See generally* Niva Elkin-Koren & Maayan Perel, *Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law*, *in* THE OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY 669 (Giancarlo Frosio ed., 2020).

241. Jean-Christophe Plantin et al., *Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook*, 20 NEW MEDIA & SOC'Y 293, 297 (2018).

242. ENGSTROM & FEAMSTER, *supra* note 222, at 16-17.

243. *Your YouTube Content & Restricted Mode*, YOUTUBE HELP, https://support.google.com/youtube/answer/7354993?hl=en [https://perma.cc/QCR7-HRTZ] (last visited Sept. 23, 2023).

244. *See, e.g.*, *Introducing New Authenticity Measures on Instagram*, INSTAGRAM (Aug. 13, 2020), https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram [https://perma.cc/LX35-9L9X].

245. *See* Mehta, *supra* note 212.

246. *See* Trendacosta, *supra* note 213.

247. Julia Alexander, *YouTube Can Now Warn Creators About Copyright Issues Before Videos Are Posted*, VERGE (Mar. 17, 2021), https://www.theverge.com/2021/3/17/22335728/youtube-checks-monetization-copyright-claim-dispute-tool [https://perma.cc/3JTY-JPJG].

with YouTube's advertising guidelines.[248] AI systems of speech moderation, hence, do not simply manage online traffic, determining which pieces of content become unavailable; they also yield regulatory consequences, directing users' behavior by sanctioning particular content and also conveying a normative message—deciding which content is deemed illegitimate. Furthermore, the recursive nature of the AI decisionmaking process could lead to scenarios where decisions on the legitimacy of specific content affect subsequent content treated by that system, giving the feedback loop of ML systems (past dependency). For instance, if the system classifies content *A* as infringing and content *B* is similar to content *A*, then content *B* is more likely to be removed (followed by content *C* and *D* and *E*, etc.)—even if the decision regarding content *A* is not in fact justifiable.

In sum, ML algorithms used in content moderation enforce speech norms and shape the behaviors and expectations of users. ML algorithms define the scope of permissible speech in a non-explicit manner by creating speech affordances, determining what content becomes available, what remains available, and to whom. These norms are driven by the economic interests of private businesses,[249] yet they constitute the digital public sphere.[250] Therefore, to overcome this democratic deficit and acquire legitimacy, they should manifest social deliberation and public participation. Accordingly, next we turn to question whether speech governance by AI could sustain the important features of democratic contestation, which are embedded in speech governance by legal norms.

IV. (THE LACK OF) CONTESTATION IN SPEECH GOVERNANCE BY AI

The transition to AI in speech moderation by platforms is not simply technical, but rather transforms the nature of speech governance. It lacks some key features that are necessary to enable society to deliberatively decide self-governing norms. Democratic contestation seeks to enable citizens to form, collectively, public opinion by facilitating discursive interactions.[251] In the following discussion, we explain why democratic contestation withers under the current system design of speech governance by AI.

A. Concentration of Rulemaking Power

One important feature of democratic contestation facilitated by law is the dispersed power to decide and interpret norms held by competing institutions and diverse human decisionmakers. By

248. *Id.*

249. Plantin et al., *supra* note 241.

250. Zittrain, *supra* note 14.

251. *See supra* Section I.B.

contrast, in speech governance by AI, systems act simultaneously as legislatures, judges, and executors when they define the classifiers, apply them to any given piece of content, and generate an outcome: whether to allow or ban it.[252] Consider YouTube's Content ID mentioned earlier as an example.[253] As noted, the system enables YouTube to automatically screen user-uploaded content and identify copyrighted material using a digital identifying code. It also determines what specific level of similarity between an uploaded video and an original copyrighted work is needed to trigger the matching feature, which will then submit a signal to the right holder, allowing her to choose whether to remove, monetize, block, or disable the allegedly infringing material before it becomes publicly available.[254]

YouTube effectively exercises judicial power when it determines which content constitutes an infringement of an original copyrighted work. It also exercises executive power when it acts to remove, disable, or filter such content. In effect, the copyright norms that govern video sharing through YouTube are shaped almost exclusively by Content ID and the data feeding it.[255] Formally, YouTube distinguishes between copyright enforcement through its Digital Millennium Copyright Act[256] compatible notice-and-takedown system and its Content ID business feature.[257] Yet essentially, considering the pervasiveness of removals though Content ID,[258] it practically redefines the meaning of copyright law in a way that solely reflects YouTube's internal business interests, leaving no room for a meaningful dialogue with other, external normative systems.

B. Diminishing Multiplicity of Meanings in Speech Norms

An important feature of speech governance by law, as discussed in Part II, is to enable individuals and groups to contest meanings on a shared ground, towards collectively deciding conflicting views, while often disagreeing. Yet the tackling of unprotected speech using AI currently fails to facilitate the same multiplicity of meanings.

---

252. *See* Niva Elkin-Koren & Maayan Perel, *Algorithmic Governance by Online Intermediaries*, *in* THE OXFORD HANDBOOK OF INSTITUTIONS OF INTERNATIONAL ECONOMIC GOVERNANCE AND MARKET REGULATION 3 (Eric Brousseau, Jean-Michel Glachant & Jérôme Sgard eds., 2019).

253. *See* Trendacosta, *supra* note 213 and accompanying text.

254. *See id.*; Perel & Elkin-Koren, *supra* note 15, at 477-78, 481, 510.

255. According to the first copyright transparency report published by YouTube in December 2021, 99% of the copyright actions processed by YouTube during the first half of 2021 were processed via Content ID. *See Access for All, A Balanced Ecosystem, And Powerful Tools*, YOUTUBE OFF. BLOG (Dec. 6, 2022), https://blog.youtube/news-and-events/access-all-balanced-ecosystem-and-powerful-tools/ [https://perma.cc/42L9-MUG4].

256. Digital Millennium Copyright Act of 1998, 17 U.S.C. § 512 (2012).

257. *What is a Copyright Claim?*, YOUTUBE HELP, https://support.google.com/youtube/answer/7002106?hl=en [https://perma.cc/6KEH-XK9J] (last visited Sept. 23, 2023).

258. *Id.*

Speech governance by AI applies data analytics techniques to identify patterns and correlations in order to classify content as unwarranted.[259] ML algorithms make predictions based on previous classifications of similar data.[260] In this respect, ML systems also differ from rule-based algorithms, which apply explicit coded definitions to particular input data ("if x then y") to generate an outcome. As noted, the input of ML algorithms is labeled data (e.g., inciting/non-inciting), which is used to train the model.[261] During the training, the algorithm will try multiple predictive rules, namely some useful correlations between multiple features and an outcome ("inciting content") and will ultimately discover which rules optimize the objective function.

ML receives outcomes (labeled data) and data as input to generate rules. Speech governance by law, to the contrary, begins with an explicit legal definition of unwarranted content and applies it to particular facts to reach an outcome (warranted/unwarranted). The Chart below illustrates these differences.

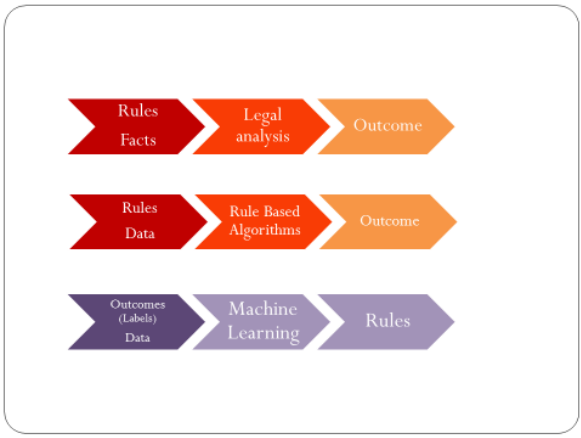

Chart 1: Rules generated by law and ML

As noted, AI classifications or predictions in content moderation systems are often followed by an operational outcome: degrading, automatic filtering, or removal of the content.[262] Often times, this is a one-shot, binary determination that either allows the content or bans it. The AI system does not engage in weighing values such as free speech and public safety or any other normative deliberation regarding the appropriate balance between them. Instead, it applies the ex ante tradeoff which it was designed to promote: if any piece of content, regardless of the speaker or the surrounding circumstances, matches

259. Gorwa et al., *supra* note 170.

260. Jonathan Zittrain, *Intellectual Debt: With Great Power Comes Great Ignorance*, MEDIUM (July 24, 2019), https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c [https://perma.cc/7JY8-NVW6].

261. *See supra* notes 240-46 and accompanying text.

262. *See supra* Section I.B.

the classifier, it shall be removed. As further elaborated below, this process is opaque, so the public has no access to the tradeoffs, values, and reasoning at stake.

C. Shrinking the Shared Ground for Public Scrutiny

There are several reasons why speech governance by AI is less susceptible to public scrutiny than speech governance by law. First, norms generated by AI are opaque and thus not subject to public scrutiny, negotiation, or social change. In contrast to the evolvement of legal norms, which relies on explicit and transparent definitions of illegal content (e.g., infringing materials or violent speech) that are subject to interpretation via processes that are open to the public, ML algorithms are designed to identify patterns and make predictions without having to explicitly reveal the norms being applied.[263] These systems do not provide explanations of their outcomes, making it more difficult for affected parties to effectively contest their outcomes and precluding any meaningful public deliberation over the legitimacy and values such outcomes may reflect.

The Global Internet Forum to Counter Terrorism (GIFCT),[264] for example, is based on a Shared Industry Hash Database (SIHD), which is kept secret to prevent gaming by adversaries.[265] Originally established in 2017 by a group of four tech firms (Facebook, Twitter, Microsoft, and YouTube), it promotes and advances the use of AI to filter terrorist propaganda by detecting images and videos that match a privately held, secretive database of content hashtags—unique digital fingerprints of alleged terrorist content, including images, videos, audio, and text.[266] This system lacks transparency even in its most fundamental component, the definition of terrorism—thus preventing parties from contesting their inclusion on a "prohibited content" list. Note that the database maintained by GIFCT is now used by thirteen different companies, including Instagram, LinkedIn, Reddit, and Snap.[267]

This is especially worrying considering the blurry boundaries between legitimate activism and illegal terrorism, which are often a matter of deep public and legal debate. A designation of a social group as a terrorist organization may critically affect its ability to operate,

263. *See supra* note 15 and accompanying text.

264. Gorwa et al., *supra* note 170; Douek, *supra* note 5; Brian Fishman, *Crossroads: Counter-Terrorism and the Internet*, 2 TEX. NAT'L SEC. REV. 82, 95-96, 97 (2019).

265. *See Technical Products*, GIFCT, https://gifct.org/tech [https://perma.cc/S56X-AGA3] (last visited Sept. 23, 2023).

266. Gorwa et al., *supra* note 170, at 2.

267. *See Story*, GIFCT, https://gifct.org/about/story/ [https://perma.cc/6NNP-7RN7] (last visited Sept. 23, 2023).

with serious legal, financial, and political consequences. Recent examples include the algorithmic targeting of Black and Muslim activist organizations.[268]

Second, AI systems conceal the value tradeoffs embedded in their optimizing function.[269] They only reveal their outcomes (e.g., classifications of content as either illegal or not), either in particular instances or in the aggregate,[270] without disclosing the meaning of the speech norm, which may involve a spectrum dependent on latent variables. Consequently, the automated classification conceals an important point of social choice on whether and how to adjust its speech norms: either by exempting particular speech or by extending the norm to cover new types of speech or circumstances. Indeed, judges too decide ad hoc and ex post which variables will be given particular weight and how. Yet, unlike judges who are explicitly required by law to state the norm and the reason of particular tradeoffs, speech norms generated by ML systems remain opaque.

Consider, for instance, the automated removal of material identified as copyright infringing. The system's designers are required to set a quantitative threshold for infringement, such as a 100% similarity, or a continuous variable indicating similarity in different samples of a music composition, or any other measure that accumulates different types of detected similarity against a certain threshold.[271] Measuring infringement simply based on the amount of identical content (e.g., a threshold number of seconds) reflects a narrow understanding of a far more elaborated legal definition of *substantial similarity*. Such a technical definition necessarily incorporates value tradeoffs, manifested by excluding some features or tweaking the system to prefer one outcome over another. Importantly, such measures, which reflect a normative judgment, are not legible to the public.

YouTube's Content ID, for instance, flagged an educational video posted by an NYU law professor that depicted a panel of experts in copyright law explaining how to analyze songs for similarity in cases of copyright infringement.[272] This was obviously because the specific quantity of the protected song that was used by the professor exceeded YouTube's technical threshold of substantial similarity. Nevertheless, under copyright law, such a transformative use made for legitimate

268. Elizabeth Dwoskin & Gerrit De Vynck, *Facebook's AI Treats Palestinian Activists like It Treats American Black Activists. It Blocks Them*, WASH. POST (May 28, 2021), https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/ [https://perma.cc/XRX4-KUA5].

269. Lehr & Ohm, *supra* note 206, at 671.

270. ML systems need not provide only binary outcomes (yes/no, true/false). Algorithms can predict quantitative or ordinal outcomes. *See* Lehr & Ohm, *supra* note 206, at 673.

271. Perel & Elkin-Koren, *supra* note 15, at 477-78.

272. Trendacosta, *supra* note 213.

educational purposes is fair.[273] If a court was called upon to decide such a copyright infringement case, its decision would have reflected particular, ex post judgment deriving from the specific circumstances of the case. In particular, fair use considerations, which are often raised by the defendant in an adversarial process, are frequently given weight.

A third barrier to public oversight of speech norms generated by ML systems is failure to disclose the procedure by which norms are decided and revised. ML systems are encumbered by previous learning, even when encountering a new and unpredictable clash of values. Thus, norms are set through a dynamic aggregation and analysis of previous instances and patterns, with the processes leading to the outcome potentially being inexplicable.

ML-based content moderation is probabilistic.[274] Decisions to ban or remove content may depend on many dynamic variables: whether the content has triggered a computational threshold, whether similar content has triggered the system before, whether third parties have flagged the content or similar content, who flagged the content, and how often these things have occurred. These variables are dynamic and opaque. They are not the result of conscious deliberation or an intentional attempt to reflect the underlying principles of our social contract. They also fail to facilitate negotiation over conflicting interests. In law, by contrast, there are explicit procedures for developing norms through judicial interpretation. Legal norms shape speech through transparent and explicit rules and standards by offering a definition of what speech is and identifying the limitations to which it is subject. While ML applications are dynamic and can shape their performance over time, the change they reflect may not necessarily be socially desirable and cannot be said to reflect a social choice. AI systems are simply not positioned to develop new conceptions of values and tradeoffs in an intelligible manner.

Finally, the variables that shape norm settings are not simply non-transparent; they can hardly be effectively scrutinized. That is due to the large scale and scope of content moderation. Currently, speech norms generated by AI could only be learned by induction from occasional instances.[275] That is the case when controversial removals

273. Another anecdote reported by a recent Electronic Foundation White Paper demonstrates how a white noise video, which is largely in the public domain, was hit by at least five strikes of YouTube's Content ID. *Id.*

274. Mike Ananny, *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance*, KNIGHT FIRST AMEN. INST. (Aug. 21, 2019), https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance [https://perma.cc/K3GX-9PB2].

275. For a proposed methodology for extracting speech norms more systematically, see Perel & Elkin-Koren, *supra* note 11.

are reported by the media[276] and when individual cases are adjudicated by courts[277] or deliberated by external bodies such as the Facebook Oversight Board.[278] Yet, these individual cases do not provide the public with sufficient information regarding the speech norms they reflect. This is especially critical given the dynamic nature of ML systems and the pace of ongoing changes in speech norms.[279]

In other cases, removal decisions are reported in bulk, in various formats of periodical transparency reports posted by digital platforms.[280] As suggested by numerous scholars, such reports are general in nature and only provide aggregated data.[281] Importantly, they typically lack specific details on the actual content that was removed and therefore fail to disclose the general norm arising from these instances.

Overall, a handful of cases, and general aggregated reports, fail to provide the public with appropriate tools to extract the actual speech norm. More importantly, they also fail to allow the democratic contestation of such norms by regulators, courts, adversarial parties, and the public at large. To facilitate meaningful contestation, it is necessary to test the speech norm against a given set of values, evaluate it, and form an opinion about their social meaning. Given the scope and scale of content moderation by AI, and given the complexity of generating norms of data, this may require new types of tools.

D. Speech Governance by AI and Democratic Contestation

Democratic mechanisms for contesting the formation of speech norms on an ongoing basis are currently lacking in AI governance of speech. In AI speech governance, the semantic nature of the law, which allows for different normative legal interpretations, is replaced by data-driven algorithms making ad hoc positive determinations right here and right now. While the learning capacities of these algorithms allow them to change their meaning over time in accordance with the data they are exposed to,[282] their dynamic decision rule is applied systematically to all cases at a given time, eliminating

276. *See, e.g.*, Taylor Hatmaker, *Facebook and Instagram Are Removing Posts Offering to Mail Abortion Pills*, TECHCRUNCH (June 28, 2022), https://techcrunch.com/2022/06/28/abortion-pills-instagram-facebook/ [https://perma.cc/F2JU-MEPS].

277. *See, e.g.*, Lenz v. Universal Music Corp., 572 F. Supp. 2d 1150, 1151-52 (N.D. Cal. 2008).

278. *See Board Decisions*, OVERSIGHT BD., https://www.oversightboard.com/decision/ [https://perma.cc/JMX2-HMWG] (last visited Sept. 23, 2023).

279. *See* Perel & Elkin-Koren, *supra* note 11, at 188-90.

280. For example, see *YouTube Community Guidelines Enforcement*, *supra* note 186.

281. *See, e.g.*, Camille François & Evelyn Douek, *The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations*, 1 J. ONLINE TR. & SAFETY 1 (2021); PASQUALE, *supra* note 15, at 18-19; Perel & Elkin-Koren, *supra* note 11, at 184-85.

282. *See supra* Section IV.C.

the possibility that different meanings might coexist. So, for instance, if a decision rule learns to target the wording "masks harmed the wearer" as misinformation,[283] its robust application might also prevent the posting of legitimate opinions and inquiries regarding mask requirements.

Table 1 summarizes the differences between governing speech by law and governing speech by AI. All in all, speech moderation by AI conceals the tradeoffs embedded in these systems, allowing critical speech norms to be extracted automatically from the labeled data, regardless of its inherent commercial biases. We currently lack sufficient tools to contest these tradeoffs and ensure they indeed reflect our social contract.[284]

In the next and final Part, we demonstrate how democratic contestation could be introduced into the design of AI systems of speech governance.

	Gov-by-Law	Gov-by-AI
Norms	Explicit	Opaque
How decided?	Deliberation, justifications	Probabilistic
Tradeoffs	Explicit, transparent	Concealed, Optimization
Adjustment, revision	Negotiated through interpretation	Shaped by data & algorithms
Ownership	Public	Private Public/private

Table 1: Governance by law and governance by AI compared

283. John W. Ayers et al., *Spread of Misinformation About Face Masks and COVID-19 by Automated Software on Facebook*, 181 JAMA INTERNAL MED. 1251, 1253 (2021).

284. *See* Langvardt, *supra* note 171.

V. SPEECH CONTESTATION BY DESIGN

A. Contestation by Design

As we have seen, speech moderation by AI is shaping public discourse without sufficient mechanisms for social contestation and deliberation. Existing systems of content moderation operate at a large scale to generate speech norms through dynamic learning shaped by data. These processes, which are opaque by nature, cannot easily be scrutinized by the public. The public lacks information not only on the type of content that has been filtered, but also on the grounds for flagging it, which makes it difficult to challenge these actions either individually or collectively. Thus, such systems undermine social dialogue and public negotiation over both the articulation and the application of speech norms. The (often efficient) mediation of disagreements over the legitimacy of content by ML systems which filter, remove, or block content comes at the cost of diminishing the democratic space for deliberating and negotiating different views on what constitutes legitimate speech and, more importantly, how to decide the scope of legitimate discourse.

Reintroducing democratic contestation into the process of implementing and crafting speech norms is therefore essential for sustaining a democratic online discourse. Accordingly, we propose incorporating *contestation by design*. The by design approach to regulation has gained prominence in protecting fundamental rights, especially privacy,[285] and more recently also in the context of content moderation.[286] Pursuant to Marco Almada, contestability by design refers to the mandatory need to build decisionmaking systems in such a way that includes the possibility to contest the outcome since their early design.[287] Much like the more established concept of privacy by design,[288] contestability by design is often described as a *design feature* meant to ensure that "human contestation of the ensuing decision will be part of its acceptance criteria."[289] Contestability by design is often focused on the individual who is subject to automated decisionmaking.[290] It is arguably derived from "the right . . . to contest"

285. *See* Ann Cavoukian, *Privacy by Design: The Definitive Workshop*, 3 IDENTITY INFO. SOC'Y 247 (2009); *see also* Hildebrandt, *supra* note 17.

286. Niva Elkin-Koren, *Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence*, BIG DATA & SOC'Y, July-Dec. 2020, at 1; Peter K. Yu, *Can Algorithms Promote Fair Use?*, 14 FIU L. REV. 329 (2020); Maxime Lambrecht, *Free Speech by Design: Algorithmic Protection of Exceptions and Limitations in the Copyright DSM Directive*, 11 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 68 (2020).

287. Almada, *supra* note 50.

288. JAAP-HENK HOEPMAN, MAKING PRIVACY BY DESIGN CONCRETE 26 (2018).

289. Almada, *supra* note 50, at 7-8.

290. Claudio Sarra, *Put Dialectics into the Machine: Protection Against Automatic-Decision-Making Through a Deeper Understanding of Contestability by Design*, 20 GLOBAL JURIST 1 (2020).

solely automated decisions under the European General Data Protection Regulation (the GDPR).[291] This right is a due process provision intended to enable people to appeal such automated decisions.[292]

However, the virtues of contestability should not necessarily be evaluated from the narrow perspective of the specific content or the individual user who suffered from an adverse decision. According to Vaccaro et al., contestability may reflect values and align the design of ML systems with contexts of use, thus promoting the perceived legitimacy of AI systems.[293] As Claudio Sarra notes:

> [T]he act of contest marks the point of transformation of the substantial juridical relationship into a more specifically procedural one. It consists in the externalized articulation of the terms of a specific dispute, which is thus made public, so that it can be articulated in a procedure that leads to a judgment.[294]

The fundamental characters of the term "contest," according to Sarra, are "publicity; the argumentative determination of the specific object to decide; [and] the transformation of the juridical relationship from substantial to procedural."[295] Therefore, contest could also mean to enable room for deliberation and dialogue over the meaning of the specific dispute's subject.[296]

As we explain next, speech contestation by design turns its focus to *public contestation*. It looks beyond the narrow interests of individual users and decisions concerning specific content and towards broader societal values concerning free speech.

---

291. *Id.*; *see* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) [hereinafter GDPR] (providing that the data controller shall implement measures to assure a "right to . . . contest the decision"); *see also* Almada, *supra* note 50; Deirdre K. Mulligan et al., *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, *in* ETHICS OF DATA AND ANALYTICS: CONCEPTS AND CASES 420 (2022).

292. *See* GDPR, *supra* note 291, art. 22(3); *see also* Kaminski & Urban, *supra* note 19.

293. KRISTEN VACCARO ET AL., CONTESTABILITY IN ALGORITHMIC SYSTEMS (2019).

294. *See* Sarra, *supra* note 290, at 7.

295. *Id.*

296. *See* TAD HIRSH ET AL., DESIGNING CONTESTABILITY: INTERACTION DESIGN, MACHINE LEARNING AND MENTAL HEALTH 95-99 (2017) (explaining that contestability is addressed via four goals and accompanying design strategies: (1) accuracy via iterative deployment and incentivizing feedback, (2) legibility by providing explanations, confidence levels, and traces of system predictions, (3) training that explicitly addresses system limitations and allows experimentation to develop shared understandings, and (4) mechanisms for questioning and disagreeing with system behavior whether at the individual or aggregate scale).

B. Speech Contestation by Design

The lack of an effective right of *individual speakers* to dispute the removal of their content has been widely addressed by the literature.[297] Enabling different stakeholders to individually dispute a specific algorithmic outcome, by individually appealing a removal decision or collectively auditing content moderation practices,[298] is important for protecting due process.[299] Nevertheless, this might be insufficient to ensure that ML content moderation systems sustain democratic contestation.[300] Since speech moderation systems do not simply provide a consumer good or service but actually mediate a social good (i.e., public discourse),[301] the absence of space for social contestation may have important social ramifications.[302] What we are lacking in speech governance by AI are procedures and processes that would enable us as a society to contest our societal speech norms. Such social contestation leaves room for disagreement while at the same time facilitates *participatory collective action*.

Importantly, the ability of *social actors* to engage in ongoing development of shared speech norms despite fundamental disagreements could reduce intolerance and societal polarization. As we noted, ML systems of content moderation are deployed to tailor content to particular users who are connected to likeminded communities.[303] Such self-reinforcing fragmentation into small tailored "publics" withers the functional utility of a democratic public sphere, as it further reduces any opportunity for platform users to be confronted with contesting views.[304] Introducing social contestation could help reinstate a "public" digital sphere in the now fragmented online speech environment.

---

297. Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 11 NEW MEDIA & SOC'Y 4366, 4378 (2018) (finding that users experience the appeal process as "speaking into a void"); Kristen Vaccaro et al., *"At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation*, *in* PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION (2022).

298. J. NATHAN MATIAS ET AL., WOMEN, ACTION, AND THE MEDIA, REPORTING, REVIEWING, AND RESPONDING TO HARASSMENT ON TWITTER (2015).

299. Jennifer M. Urban & Laura Quilter, *Efficient Process or Chilling Effects: Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUT. & HIGH TECH. L.J. 621, 626 (2006).

300. *See* Trendacosta, *supra* note 213; *see also* Sarra, *supra* note 290 (explaining that contestability is distinct from simply opposing the outcome of ML; rather, it is the ability to engage with the substance of the decision itself).

301. Candace Cummins Gauthier, *Right to Know, Press Freedom, Public Discourse*, 14 J. MASS MEDIA ETHICS 197 (1999).

302. Sarra, *supra* note 290, at 1 ("A more general concern can be raised about the social opportunity to let significant sectors of social life be guided completely by machines.").

303. *See supra* notes 74-78 and accompanying text.

304. Anat Ben-David, *Counter-Archiving Facebook*, 35 EUR. J. COMMC'N 249, 255-56 (2020). *See generally* KATHLEEN HALL JAMIESON & JOSEPH N. CAPPELLA, ECHO CHAMBER: RUSH LIMBAUGH AND THE CONSERVATIVE MEDIA (2008).

Faced with a robust system of algorithmic speech regulation, which is generating and exacerbating inscrutable data driven speech norms, could democratic procedures for contestation be embedded in the design of this system? As we show next, we believe they can.

C. Embedding Speech Contestation by Design

How can one restore democratic contestation in speech governance by AI? There are different approaches to incorporating contestability in ML systems.[305] Some approaches rely on adding non-functional software requirements to ensure that users can have the necessary data and tools to exercise their right to contest.[306] Building explainable ML systems is one example.[307] However, while these approaches are important to provide users with explanations regarding the automated decision affecting them, or to assure the ML system works as intended,[308] they seem to focus on embracing the narrow rights of the affected user rather than promoting broader social interests.

A different approach for designing contestable ML systems is based on participatory design.[309] It proposes to incorporate, at each stage of the development of the ML system, feedback from relevant stakeholders that might be affected by the system.[310] Rather than focusing on facilitating intervention by a specific user, this approach may provide a form of collective ex ante and ongoing intervention in the processing of the ML system.[311] Mireille Hildebrandt, for instance, proposes "agonistic machine learning," namely, "demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modeling the same event, person[,] or action."[312] Such built-in falsifiability, she argues, could ensure that "those who will suffer or enjoy the consequences are heard and their points of view taken into account."[313] In a similar vein, Finn Brunton and Helen Nissenbaum argue that obfuscation could be applied as a strategy for contesting data collection

305. *See* Almada, *supra* note 50, at 10.

306. *Id.*

307. *See generally* Maja Brkan, *Do Algorithms Rule the World?: Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond*, 27 INT'L J.L. & INFO. TECH. 91, 92 (2019).

308. *See generally* Joshua Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

309. Janet Davis, *Design Methods for Ethical Persuasive Computing*, *in* PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON PERSUASIVE TECHNOLOGY 1 (2009).

310. IAN SOMMERVILLE, SOFTWARE ENGINEERING (Marcia Horton et al. eds., 9th ed. 2011).

311. Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SEC. & PRIV. 46, 46 (2018).

312. Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agnostic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 106 (2019).

313. *Id.* at 109.

techniques.[314] Obfuscation protects privacy—i.e., makes it harder to infer sensitive information—by introducing more "noise" into the data, disrupting the collection, aggregation, and processing of data by blurring signal and noise.[315] Others, such as Kulynych et al., have proposed subversive strategies that would enable stakeholders to counter optimizing systems from the outside.[316]

Another option to build contestable ML systems is by automating the process of reviewing the automated decision.[317] Embedding a trusted third party algorithm in the design of the ML system could achieve this purpose.[318] More importantly, as demonstrated below, it could be specifically useful for the purpose of giving voice to divergent conceptions of free speech, while formulating space for deliberation and contestation at the processing stage, prior to the automatic production of speech norms.

Our proposal for *speech contestability by design* seeks to promote processes and procedures that introduce contestation into content moderation systems. This approach is inspired by the contestation processes and procedures that are embedded in the law.[319] As we noted, legal norms enable social actors to agree on high-level principles and work out the details of the required tradeoffs as courts apply these principles to particular cases down the road.[320] This enables disagreement while keeping society whole. Yet speech contestability in AI speech governance cannot be achieved by relying exclusively on legal principles and judicial review. The scale and robustness of algorithmic speech moderation and the dynamic nature of AI systems suggest that a supplementary design approach is necessary in order to effectively facilitate contestation in speech governance by AI.

---

314. *See generally* FINN BRUNTON & HELEN NISSENBAUM, OBFUSCATION: A USER'S GUIDE FOR PRIVACY AND PROTEST (2016).

315. *Id.*

316. Bogdan Kulynych et al., *POTs: Protective Optimization Technologies*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2020).

317. Almada, *supra* note 50, at 10.

318. *Id.*

319. Introducing contestation into AI-based speech governance could be pursued through other technological means that rely on the unique attributes of governance by AI without necessarily replicating the pluralizing mechanisms of the law. Ellen Goodman, for instance, argues that "[d]igital enclosure seals communicators in feedback loops of data that are harvested from attention and then used to deliver content back to data subjects in an endless scroll." She describes several technological forms of friction that may be applied to disrupt the viral spread of disinformation. For instance, communication delays that encourage speakers to think before they publish, along the lines of "are you sure you want to say X?", could be systematically embedded in AI-based systems of speech governance. Another form of technological friction described by Goodman is "virality disruptors"—technologies employed "to disrupt traffic at a certain threshold of circulation," so as to reduce the salience of low-fidelity communication. The sharing limit imposed by WhatsApp offers one example of such a disrupter. Ellen P. Goodman, *supra* note 38, at 648, 651.

320. *See supra* notes 106-18 and accompanying text.

Next, we briefly demonstrate how ML content moderation systems could be reconfigured in order to advance the democratic notion of contestation in managing online public discourse.

1. Embedding an Adversarial Approach

One way to promote participatory public engagement in setting speech norms by AI systems is to incorporate adversarial procedures in the system design. An adversarial approach, inspired by law, could guide the creation of contesting algorithms, which would automate the process of contesting decisions about speech.[321]

Adversarial legal procedures, where parties are called to present their contesting positions in front of a judge or jury, are among the fundamental elements of common law justice systems and the gold standard of dispute resolution.[322] The underlying assumption of adversarial procedures is that laying out the contesting positions in a dispute is the best way to test factual evidence and reach sound decisions.[323] In ML, adversarial learning algorithms, whose goal is to identify weaknesses, are often deployed to monitor algorithmic black boxes. For instance, Generative Adversarial Networks (GANs) make use of an unsupervised ML to automatically identify learning and patterns in the main ML system.[324]

Content moderation by ML currently lacks comparable adversarial mechanisms. Consequently, a system that is designed to optimize a functional objective, such as removing any materials which match sampled content provided by copyright holders, is likely to overlook a wide range of social interests that might be implicated by this choice, such as enabling educational use of copyrighted material[325] or protecting political parody.[326]

Contesting algorithms offer one way to introduce an adversarial feature into ML content moderation systems. Under this proposal, any content subject to removal would have to be run through a competing system designed to reflect a declared set of societal values.[327] This virtual checkpoint, or "Public AI Content Moderation System," as we propose to call it, would algorithmically judge the content that was flagged for removal against norms generated dynamically by independent bodies, such as public civil society organizations or the

321. Elkin-Koren, *supra* note 286.

322. Ellen E. Sward, *Values, Ideology, and the Evolution of the Adversary System*, 64 IND. L.J. 301, 312 (1989).

323. *Id.* at 302.

324. *See, e.g.*, JOST TOBIAS SPRINGENBERG, UNSUPERVISED AND SEMI-SUPERVISED LEARNING WITH CATEGORICAL GENERATIVE ADVERSARIAL NETWORKS (2016).

325. *See, e.g.*, Ann Bartow, *Educational Fair Use in Copyright: Reclaiming the Right to Photocopy Freely*, 60 U. PITT. L. REV. 149 (1998).

326. *See, e.g.*, Cathay Y.N. Smith, *Political Fair Use*, 62 WM. & MARY L. REV. 2003 (2021).

327. Elkin-Koren, *supra* note 286, at 10.

judiciary.[328] Should the contesting algorithm reaffirm the platform's removal decision, removal would proceed. Should the contesting algorithm reach a different conclusion, removal of the content would be postponed until the conflict is resolved, either algorithmically, based on scorings produced by the two systems, or through human review. Following resolution, both systems would be updated with the results of the deliberation.

Contesting algorithms would be guided by several main principles. First, the adversarial system would add a separate independent layer to the dominant system of content moderation, rather than attempting to reconfigure the dominant system and its optimization model. Second, the adversarial model would be dynamic and updated in accordance with evolving norms.[329] Third, the adversarial model would seek to disclose controversy, rather than mandating any particular social tradeoff.[330] More generally, the proposed adversarial design would establish two independent automated processes for deciding values—one private and at least one public. Each system could be designed to optimize different objectives, potentially reflecting different tradeoffs. Contestation could reveal these tradeoffs and make them explicit, subject to open deliberation and public scrutiny. Consequently, such design intervention may facilitate a common ground for public deliberation over social choices regarding speech norms.

Creating a legal duty, or otherwise providing incentives to automatically check removable content on an independent external system, may contribute to democratic contestability in several ways. First, contesting algorithms offers an effective method for acquiring information on specific removal decisions, at scale. As discussed above, transparency reports published periodically by digital platforms fail to disclose the general speech norm arising from the accumulated instances.[331] The ability to effectively review each of these removal instances algorithmically provides an opportunity to establish more knowledge not only on the general scope and scale, but also on the substantive choices involved in the platform's removal policy.

---

328. A virtual public check point might also consist of several systems, each advancing the legibility of speech norms generated by the platform's AI moderation system—by testing its outcome against an algorithm reflecting a distinct set of tradeoffs.

329. Elkin-Koren, *supra* note 286.

330. Mandating particular tradeoffs on platforms by law might be considered a radical intervention in freedom of speech, and as such could be judged unconstitutional. *See* Keller, *supra* note 11, at 13. However, the adversarial approach envisioned here takes a procedural approach, which does not prioritize any particular speech norm; it merely creates a mechanism for contesting competing values in an algorithmic environment. It would not seek to mandate platforms to embed any particular social tradeoffs in applying content moderation by AI, something that would require an ex ante definition of the scope of free speech.

331. *See YouTube Community Guidelines Enforcement*, *supra* note 186; *see also* François & Douek, *supra* note 281.

Second, algorithmic contestation could help extract the speech norms embedded in AI systems of content moderation and to test them against alternative norms. More precisely, content moderation always involves tradeoffs between conflicting values and interests of various stakeholders. The monolithic design of content moderation by AI makes it difficult to clearly identify the political choices that underlie decisions made by existing systems, such as which features were considered in determining legitimate use and the weight given to each.[332] The adversarial procedure would force the disclosure of necessary information on removal decisions that may be inconsistent with some social values, thus enabling judicial review or public oversight of such gaps. Contestability might turn the inscrutable outcome of AI systems to be more amenable to normative reasoning. Thus, in addition to serving as an external check over platforms' non-transparent content moderation practices, applying contesting algorithms could facilitate judicial and public deliberation over competing values. Over time, it may also help to align the robust content moderation systems of digital platforms with a more diverse set of speech norms.

Third, algorithmic contestation moves beyond transparency reports and disclosure duties to facilitate an ongoing form of public engagement with speech norms generated by AI. Algorithmic contestability could effectively counteract the feedback loop of content moderation systems run by platforms, which are set up to optimize a predetermined tradeoff reflecting their business interests. Such a procedure could create an ongoing and dynamic *check* as well as *counter pressure* against platforms' monolithic content removal systems.

The democratic governance structure leaves room for disagreement by creating institutions where tradeoffs can be deliberated, negotiated, and decided (elections) and where they are subject to oversight (judicial review). To preserve such pluralism in ML content moderation systems, the adversarial strategy takes a procedural approach. It does not set any particular norm, but instead creates a procedure for contesting competing values in an algorithmic environment. This is a democratic move: we do not need to reach consensus on the tradeoffs, but instead can agree on a legitimate procedure by which these tradeoffs can be decided. Moreover, articulating the values and interests of stakeholders that are underrepresented by the dominant AI removal system creates a space for developing a comprehensive public alternative to that system. In this context, AI governance offers new opportunities since the aggregation of individual models, and the resulting policy operations of different stakeholders, are digitally coded.

332. *See supra* notes 237-51 and accompanying text.

All in all, the adversarial design creates a common ground for negotiating speech norms—a procedural framework under which competing tradeoffs are confronted to produce an outcome that actually reflects social negotiation. It enables public scrutiny over speech moderation, thus facilitating a more democratic evolvement of online speech norms.

2. Separation of Functions

Another possible means to promote democratic contestation by design is to inject the legal principle of separation of powers into algorithmic content moderation through *separation of functions*. The idea is to separate different functions performed by the monolithic AI content moderation systems of digital platforms and to outsource the law enforcement functions to external, independent, unbiased algorithms.[333] Currently, the public law enforcement functions of social media platforms are integrated with private business functions that are driven by commercial interests.[334] The same technical design that is used for targeted advertising and for curating personalized content is also deployed for monitoring and enforcing speech norms. The system is informed by the same labeling of users and content and makes use of the same application programming interfaces, learning patterns, and software. *Separation of functions* would stimulate the creation of alternative solutions to content moderation, thereby supporting the development of more diversified speech norms in AI systems.

This approach is different from contesting algorithms in two main ways. First, the two approaches differ in the space in which the deliberation between competing values occurs. *Contesting algorithms* would add a layer to the existing system, creating a separate process that would essentially subject the platforms' private content removal decisions to public review. By contrast, *separation of functions* proposes to enable external deliberation executed on independent grounds without affecting the platforms' private content moderation systems: the two systems would simply be concerned with different tasks. Second, while the friction advocated by *contesting algorithms* might be triggered by each and every removal choice, under *separation of functions*, friction applies only to removals that are based on the platforms' legal duties. This solution—like contesting algorithms—also offers an alternative to reconfiguring the original AI-based system of content moderation and attempting to alter the optimization model. Therefore, it sustains a distinction between the rights and duties of private actors and their public functions. However, it does so by creating a separate and independent public system to flag and remove

333. Elkin-Koren & Perel, *supra* note 174, at 893.

334. *Id.*

unlawful content (i.e., unwarranted content as defined by the law). This could encourage platforms to keep their systems' commercial functions distinct from their law enforcement functions to ensure that the proposed external law enforcement system does not disrupt their business interests.

CONCLUSION

Democracy depends on a functioning framework for negotiating differences, adjusting positions, modifying opinions, and making concessions. Since AI systems have become the go-to architecture for moderating public discourse, it is now essential to enable contestation in their design, which would better reflect collective social choice. This Article called to reflect the value perceptions of diverse stakeholders in AI-based systems of content moderation through bottom-up strategies. Giving room to different perceptions of values as held by different members of society could decentralize the tremendous power of platforms to decide tradeoffs in speech regulation in a non-transparent way. Rather than having these tradeoffs determined unilaterally by the platform or by top-down regulation, they would be shaped by society while also injecting diversity and securing contestation in the algorithmic governance of speech.

Platforms should promote an infrastructure that would facilitate ongoing public engagement with speech norms. In some cases, this may involve securing independent access to content moderation outcomes to enable algorithmic contestation. In other instances, it may require platforms to enable access to public algorithms that would perform law enforcement functions. The challenge to policymakers would be to encourage the development of technological designs as well as social institutions that can reinstate the virtues of contestation in online flows moderated by ML.[335]

335. One possible way to encourage the application and use of speech contestability by design is by imposing legal duties on users of social media. For instance, in a recent decision, the Israeli Supreme Court held that sharing (but not simply "liking") a defamatory post on social media may constitute a "publication" subject to liability under Israel's Anti-Defamation Law. *See* CivA 1239/19 Shaul v. Nidaily Communications, Ltd., (2020) (Isr.). Addressing a similar challenge, though reaching a different conclusion, the Supreme Court of Switzerland recently held that users could be held liable for "likes" and "shares" of defamatory posts on Facebook. A legal duty to think before you share may slow down potentially viral and automatic dissemination of content while encouraging users to consult their inner voice and its moral code more often. In such cases, the friction is the human in the loop who exercises his own conception of morality, effectively contesting the automatic design of speech norms with his own. *See* Bundesgericht [BGer] [Federal Supreme Court] Jan. 29, 2020, 6B_1114/2018 (Switz.). Alternatively, it is possible to put legal pressure on platforms (such as conditioning platform immunity on actions that introduce ways of deliberating and contesting speech norms) or economic pressure (such as imposing fines on platforms that fail to do so). Goodman, *supra* note 38.