Supporting Democracy through Content-Neutral Social Media Policies

Christopher L. Quarles
University of Michigan, School of Information, Ann Arbor, Michigan

DOI hyperlink: https://doi.org/10.38126/JSPG220108

Corresponding author: chrisquarles@gmail.com

Keywords: Social Media Policy; content moderation; information infrastructure; de-amplification

Executive Summary: The internet and social media carry vast amounts of new information every second. To make these flows manageable, platforms engage in content moderation, using algorithms and humans to decide which content to recommend and which to remove. These decisions have profound effects on our elections, democratic debate, and human well-being. The U.S. government cannot directly regulate these decisions due to the scale of the content and the First Amendment. Rather than focusing exclusively on whether or what content gets moderated, policy-makers should focus on ensuring that incentives and processes create an information infrastructure that can support a robust democracy. These policies are most likely to be content-neutral. Three content-neutral mechanisms are promising targets for policy: process, transparency, and de-amplification.

I. Introduction

Online spaces have become the modern "public sphere", where new ideas are put forth and explored in a model of democratic discourse (Habermas 1991). However, the information infrastructure of the internet has caused this discourse to become pathological in many ways. Market segmentation has created conflicting streams of information, which causes different groups to have inconsistent understanding of the truth (Iyengar and Massey 2019). Politicians strategically become less civil to garner support on social media (Frimer et al. 2022). Foreign agents can more easily undermine our confidence in elections (Jensen 2018). And unchecked toxicity causes reasonable people to disengage from online spaces (Bail 2021). The causes of these problems are fundamentally new to humanity. Exabytes of new data are created every day, leading to immense challenges in online moderation.

i. Trustees of the Public Sphere

Social media platforms are the trustees of the modern public sphere. They are currently the only entities with both the access and resources to moderate the discussion essential for a functioning democracy (Gillespie 2018). However, their primary incentive is profit, not meaningful debate. While these goals often align, there have been notable conflicts such as the Cambridge Analytica scandal (Klonick 2020).

Unhappy with platform moderation, some state legislatures have passed bills forbidding social media bans of politicians or political speech (Vidales 2022). In response, courts have cited corporate speech rights, claiming that online moderation of user views is an exercise of social media platforms' constitutional rights (Zakrzewski 2022). Neither of these approaches will support democracy in the long term. Absent moderation, online spaces become toxic, which pushes away many users (Gillespie 2018). However, privileging corporate speech over human speech creates precedent for dystopian corporate censorship. Policies must strike a balance between the free flow of dialogue and the creation of friendly spaces for that dialogue to happen.

Americans have debated the boundaries of free speech since the First Amendment was drafted, and these conversations will continue. As norms and technology change, so too does the content we consider acceptable. Certainly, there are types of content, such as child sexual abuse material, where legislation limiting the free flow of information is warranted (The United States Department of Justice).

ii. Information Infrastructure

Policy makers should differentiate policies that directly address online *content* from those that address *information infrastructure.* Information infrastructure is the interacting system of technology, government & corporate policies, and human behavior, insofar as they affect how information is transmitted in society. Notably, it is embedded in code, norms, and human psychology as well as in physical infrastructure. Early American information infrastructure consisted primarily of face-to-face conversations and letters (Burke 2005), as well as mass communication in the form of pamphlets, newspapers, and sermons (Harshman, Hill, and Moran 2020). For the Founding Fathers, free speech referred to the unhindered production of verbal dialogue, letters and pamphlets (Bogen 1983).

With millions of people posting on the internet each day, *production* of speech is no longer an issue. Instead, the important decisions are around *consumption* and *filtering*. There is no question about whether platforms should moderate content – they must (Gillespie 2018). The free flow of speech is impossible, since the number of posts is far more than any user's brain can handle. To moderate this firehose of content, social media corporations typically use multiple complex layers of machine learning algorithms and human moderation (Gillespie 2018; Lada, Wang, and Yan 2021). Instead of trying to control that moderation directly, policy-makers should set up incentive structures to ensure the information infrastructure preserves democratic values.

Content-neutral policies, which do not single out specific subject matter are promising for this purpose, because (a) they are particularly well-adapted to addressing systemic issues, (b) they are less likely to become politicized due to their lack of focus on specific content, and (c) they are more likely to hold up to First Amendment scrutiny (Keller 2021). Many promising content-neutral categories either focus on ensuring that platforms have adequate processes, promote transparency, or require de-amplification of the most viral information.

II. Process-Based Policies

As trustees of the public sphere, social media platforms have the responsibility to maintain standards for democratic dialogue. While the government might be limited in its ability to regulate specific moderation decisions, it can promote decision-making that is consistent with a well-functioning information infrastructure. Process-based policies can ensure that moderation is consistent and answerable to the people.

i. Information Fiduciaries

One promising approach is to treat platforms as *information fiduciaries* (Balkin 2016; Rhum 2021). A financial consultant is required to act in their customers' best interests, and a doctor's speech rights are limited when it comes to patient privacy (Balkin 2018). Similarly, platforms could be expected to provide due process around content moderation, safeguard personal data, or keep moderation decisions functionally separate from profit-motivated decisions.

The motivation for this comes from the sensitive position of users. Users have neither the ability to understand how their actions are being manipulated, nor control of the data that platforms collect. Furthermore, while there are multiple social media platforms, they are not substitutable. The purpose and user base of a platform make it hard for users to switch. Their market dominance, hidden algorithms, and data collection puts platforms in a privileged position, so it is reasonable for the government to expect a fiduciary responsibility. Fiduciary obligations should extend to any company that traffics in end-user content or data, such as ad servers and subcontractors.

One benefit of the information fiduciary approach is that the underlying responsibility is clear, regardless of changing circumstances. Platforms vary in the way they moderate, share user data, and earn money. As new technologies bring new modes of operating, the idea of an information fiduciary will last beyond any particular technology.

ii. Public oversight

Platforms could also be required to implement public oversight of moderation decisions. Public oversight models that could be adapted to policy already exist. Meta's Oversight Board uses an expert panel to selectively review moderation issues, much like the U.S. Supreme Court (Klonick 2020). Reddit uses a large network of unpaid community moderators. Since platforms are trustees of the public sphere, it is reasonable to expect oversight of platform decisions about moderation to ultimately rest with the people. Any such legislation should set clear procedural expectations for participation and require the oversight to have binding authority on the platforms' decisions. The supervisory body could have oversight over the Terms of Service (TOS), over the platforms' decision-making subject to their TOS, or, as in the case of Meta's Oversight Board, both.

However, this approach is problematic in that it hands American information infrastructure to extra-judicial courts. U.S. law has evolved over the past 250 years. It is not clear that we can expect independent oversight bodies to consistently adjudicate wisely, nor does public oversight address the problem of political fragmentation. Competing, politically aligned platforms could conceivably create oversight processes that reinforce or even magnify their partisan leanings.

III. Transparency Policies

Given the importance of online spaces for public discourse and their influence on perceptions of public opinion, it is reasonable to expect transparency in how those decisions are made (McGregor 2019). Increased transparency is a necessity for the future of the internet, as platforms have a long way to go towards increasing transparency around both moderation and internal policies. However, an expectation of complete transparency is neither practical nor technically feasible. Complete transparency makes algorithms more manipulable by those with the resources to do so, which can drown out individual users and decrease the quality of user experiences (Hosanagar and Jair 2018). Modern artificial intelligence (AI) recommender systems "learn" the types of content to promote on their own. Making these decisions completely transparent is impossible, since the software engineers who design the algorithms often do not understand all the recommendation decisions (Simeone 2018).

i. Research Transparency

Researchers have consistently called for more access to social media data (Bruns et al. 2018; Leonelli et al. 2021). Data availability often depends on whether user posts are public, with public platforms like Twitter and Reddit historically having the most accessible data (Bruns 2019). Platforms face a difficult decision in balancing user privacy and technical limitations with transparency for researchers (Walker, Mercea, and Bastos 2019). Thus, decisions about data availability can be difficult and require advanced expertise. However, platforms have repeatedly prioritized profitability and public image over other factors (Gillespie 2018).

Legislation could mandate the creation and oversight of a data sharing framework, modeled on pre-existing frameworks (Bruns et al. 2018; Nicholas and Thakur 2022; Harvard University). Day-to-day decisions on such a framework should be answerable to an external ethics board rather than to corporate management. Since moderation decisions are made by AI, which influences user behavior based on personalized information, researchers need experimental access to algorithms (Greene, Martens, and Shmueli 2022). The framework would also need to be responsive to changing technology.

ii. Bot Transparency

AI-enabled accounts have repeatedly been used to deceive the public (Ferrara et al. 2016). While the government cannot mandate that *people* tell the truth, AI does not yet have First Amendment rights (Finkel et al. 2017). Congress could pass a "bot bill" requiring AI-generated content or AI-managed accounts to be labeled, such as the Bot Disclosure and Accountability Act[1] or California's Bolstering Online Transparency Act[2]. Without legislation it is likely that no uniform standard will emerge, since deceptive use of AI can be quite profitable for marketers (Luo et al. 2019).

To be effective, a bot bill would need to require enforcement from platforms, since they have the access to detect, label, and remove bots (DiResta

[1] Bot Disclosure and Accountability Act of 2019, S.2125, 116th Cong. (2019)

[2] Bolstering Online Transparency Act, California SB-1001 (2018), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

2019). This requirement must be carefully tailored, as overzealous enforcement enables malicious reporting and removal of human accounts. AI's are used for a variety of purposes, from grammar-checking to chatbot networks. Thus, any bot bill should clearly define what qualifies as a bot. The policy should address platforms' own use of AI to recommend content, by requiring plain-language summaries of the data used to make decisions and how the data was acquired.

IV. De-amplification Policies

The internet, combined with our own psychological tendencies, amplifies outrage and extreme points of view (Brady et al. 2021), which can magnify misinformation (Carrasco-Farré 2022) and political extremism (Hasell 2021), suppress meaningful democratic dialogue (Hampton et al. 2014), and create misperceptions of public opinion (Quarles and Bozarth 2022). While amplification can be caused by algorithms, it happens naturally due to the online information infrastructure. Unlike traditional forms of communication that require more effort, online content is reshared quickly and often reflexively. This natural amplification process, called preferential attachment, naturally leads to some signals being orders of magnitude more popular than average (Barabási and Albert 1999).

i. Circuit-breakers & dampers

Policies requiring "friction" have the potential to temper amplification. Scholars have proposed "circuit-breakers" which would temporarily stop algorithmic amplification when a piece of content is spreading too fast (Goodman 2021; Keller 2021). Human moderators could then manually examine the content, which may be useful for certain cases like misinformation (Simpson and Conner 2020). A more content-neutral approach is a "damper". Rather than having a single discrete point where a human moderator steps in, a damper would algorithmically decrease the probability that a post is recommended gradually as it spreads more quickly on the network. A policy requiring dampers would limit the spread of viral information and make the information infrastructure more representative of face-to-face human communication. Unfortunately, there is little research on friction-based de-amplification, and its systemic effects on social media are not well understood.

V. Conclusion

We are in the midst of a major transition in humanity's ability to communicate. While providing amazing benefits, the internet has amplified extremism and incivility, decreased information diversity, and allowed a small number of people to limit public conversations. However, these issues are merely symptoms of a highly-connected information infrastructure that has evolved to support corporate profits and Silicon Valley culture. By crafting thoughtful, content-neutral social media policies, the government can create an information infrastructure that supports deliberative democracy and, in the process, solve many of the internet's problems.

Given the scale and complexity of the problem, multiple approaches will be necessary. Some policies should be procedural, for instance by creating an information fiduciary requirement or requiring public oversight mechanisms for platforms. A framework for research transparency would allow for better decision-making in the future. And an AI disclosure bill would help separate protected human speech from self-serving manipulation by powerful interests. In addition, government should make a significant effort to de-amplify viral information, which will allow more diverse points of view. Like the economy influences how money is exchanged, so too does our information infrastructure affect how we share and exchange information. Given the importance of protecting free speech, content-neutral social media policies are an essential tool for ensuring that internet communication supports democratic values.

References

Bail, Chris. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton, NJ: Princeton University Press.

Balkin, Jack M. 2016. “Information Fiduciaries and the First Amendment.” *U.C. Davis Law Review* 49 (4): 1183–1234.

Balkin, Jack M.. 2018. “Free Speech Is a Triangle.” *Columbia Law Review* 118 (7): 2011–56.

Barabási, Albert-László, and Réka Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286 (October): 509–12. https://doi.org/10.1126/science.286.5439.509.

Bogen, David. 1983. “The Origins of Freedom of Speech and Press.” *Maryland Law Review* 42 (3): 429–65.

Brady, William J., Killian McLoughlin, Tuan N. Doan, and Molly J. Crockett. 2021. “How Social Learning Amplifies Moral Outrage Expression in Online Social Networks.” *Science Advances* 7 (33): 1–15. https://doi.org/10.1126/sciadv.abe5641.

Bruns, Axel. 2019. “After the ‘APIcalypse’: Social Media Platforms and Their Fight against Critical Scholarly Research.” *Information, Communication & Society* 22 (11): 1544–66. https://doi.org/10.1080/1369118X.2019.1637447.

Bruns, Axel, Anja Bechmann, Jean Burgess, Andrew Chadwick, Lynn Schofield Clark, and et al. 2018. “Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process.” *Internet Policy Review*. https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786.

Burke, Kathryn. 2005. “Early American Letter Writing.” https://postalmuseum.si.edu/research-articles/letter-writing-in-america.

Carrasco-Farré, Carlos. 2022. “The Fingerprints of Misinformation: How Deceptive Content Differs from Reliable Sources in Terms of Cognitive Effort and Appeal to Emotions.” *Humanities and Social Sciences Communications* 9 (1): 162. https://doi.org/10.1057/s41599-022-01174-9.

DiResta, Renee. 2019. “A New Law Makes Bots Identify Themselves - That’s the Problem.” *Wired*, July 24, 2019. https://www.wired.com/story/law-makes-bots-identify-themselves/.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. “The Rise of Social Bots.” *Communications of the ACM* 59 (7): 96–104. https://doi.org/10.1145/2818717.

Finkel, Jacob, Steven Jiang, Mufan Luo, Rebecca Mears, Danaë Metaxa-Kakavouli, Camille Peeples, Brendan Sasso, Arjun Shenoy, Vincent Sheu, and Nicolás Torres-Echeverry. 2017. “Fake News and Misinformation: The Roles of the Nation’s Digital Newsstands, Facebook, Google, Twitter and Reddit.” https://law.stanford.edu/wp-content/uploads/2017/10/Fake-News-Misinformation-FINAL-PDF.pdf.

Frimer, Jeremy A., Harinder Aujla, Matthew Feinberg, Linda J. Skitka, Karl Aquino, Johannes C. Eichstaedt, and Robb Willer. 2022. “Incivility Is Rising Among American Politicians on Twitter.” *Social Psychological and Personality Science*, April. https://doi.org/10.1177/19485506221083811.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Goodman, Ellen P. 2021. “Digital Fidelity and Friction.” *Nevada Law Journal* 21 (2): 623–54.

Greene, Travis, David Martens, and Galit Shmueli. 2022. “Barriers to Academic Data Science Research in the New Realm of Algorithmic Behaviour Modification by Digital Platforms.” *Nature Machine Intelligence* 4 (4): 323–30. https://doi.org/10.1038/s42256-022-00475-7.

Habermas, Jürgen. 1991. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press.

Hampton, Keith N, Lee Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. 2014. “Social Media and the ‘Spiral of Silence.’” Washington, DC. http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/.

Harshman, Jason, Rebecca Hill, and James Moran. 2020. "Media and Communication Technology in the Making of America." *EDSITEment!*, 2020. https://edsitement.neh.gov/closer-readings/media-and-communication-technology-making-america.

Harvard University. n.d. "Social Science One." Accessed January 9, 2023. https://socialscience.one/.

Hasell, Ariel. 2021. "Shared Emotion: The Social Amplification of Partisan News on Twitter." *Digital Journalism* 9 (8): 1085–1102. https://doi.org/10.1080/21670811.2020.1831937.

Hosanagar, Kartik, and Vivian Jair. 2018. "We Need Transparency in Algorithms, but Too Much Can Backfire." *Harvard Business Review*, July 2018. https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire.

Iyengar, Shanto, and Douglas S Massey. 2019. "Scientific Communication in a Post-Truth Society." *Proceedings of the National Academy of Sciences* 116 (16): 7656–61. https://doi.org/10.1073/pnas.1805868115.

Jensen, Michael. 2018. "Russian Trolls and Fake News: Information or Identity Logics?" *Journal of International Affairs* 71 (1.5): 115–24.

Keller, Daphne. 2021. "Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard." https://knightcolumbia.org/content/amplification-and-its-discontents.

Klonick, Kate. 2020. "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression." *Yale Law Journal* 129 (8): 2418–99.

Lada, Akos, Meihong Wang, and Tak Yan. 2021. "How Does News Feed Predict What You Want to See?" 2021. https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/.

Leonelli, Sabina, Rebecca Lovell, Benedict W. Wheeler, Lora Fleming, and Hywel Williams. 2021. "From FAIR Data to Fair Data Use: Methodological Data Fairness in Health-Related Social Media Research." *Big Data & Society* 8 (1). https://doi.org/10.1177/20539517211010310.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases." *Marketing Science* 38 (6): mksc.2019.1192. https://doi.org/10.1287/mksc.2019.1192.

McGregor, Shannon C. 2019. "Social Media as Public Opinion: How Journalists Use Social Media to Represent Public Opinion." *Journalism* 20 (8): 1070–86. https://doi.org/10.1177/1464884919845458.

Nicholas, Gabriel, and Dhanaraj Thakur. 2022. "Learning to Share: Lessons on Data-Sharing from Beyond Social Media." https://cdt.org/insights/learning-to-share-lessons-on-data-sharing-from-beyond-social-media/.

Quarles, Christopher L., and Lia Bozarth. 2022. "How the Term 'White Privilege' Affects Participation, Polarization, and Content in Online Communication." *PLOS ONE* 17 (5): e0267048. https://doi.org/10.1371/journal.pone.0267048.

Rhum, Kimberly. 2021. "Information Fiduciaries and Political Microtargeting: A Legal Framework for Regulating Political Advertising on Digital Platforms." *Northwestern University Law Review* 115 (6): 1829–73.

Simeone, Osvaldo. 2018. "A Very Brief Introduction to Machine Learning With Applications to Communication Systems." *IEEE Transactions on Cognitive Communications and Networking* 4 (4): 648–64. https://doi.org/10.1109/TCCN.2018.2881442.

Simpson, Erin, and Adam Conner. 2020. "Fighting Coronavirus Misinformation and Disinformation." https://www.americanprogress.org/article/fighting-coronavirus-misinformation-disinformation/.

The United States Department of Justice. n.d. "Citizen's Guide to U.S. Federal Law on Child Pornography." Accessed September 9, 2022. https://www.justice.gov/criminal-ceos/citizens-guide-us-federal-law-child-pornography.

Vidales, Jesus. 2022. "Texas Social Media 'Censorship' Law Goes into Effect after Federal Court Lifts Block." *The Texas Tribune*, September 16, 2022. https://www.texastribune.org/2022/09/16/texas-social-media-law/.

Walker, Shawn, Dan Mercea, and Marco Bastos. 2019. "The Disinformation Landscape and the Lockdown of Social Platforms." *Information, Communication & Society* 22 (11): 1531–43. https://doi.org/10.1080/1369118X.2019.1648536.

Zakrzewski, Cat. 2022. "11th Circuit Blocks Major Provisions of Florida's Social Media Law." *Washington Post*, May 23, 2022. https://www.washingtonpost.com/technology/2022/05/23/florida-social-media-11th-circuit-decision/.

Christopher L. Quarles is a PhD candidate in the School of Information, a researcher with the Center for Ethics, Society & Computing, and a fellow at the Stone Center for Inequality Dynamics. His current research focuses on how information technology affects how we group ourselves, and on systemic trends in inequality and opportunity. In the long term, he hopes to have a practical impact on the way our information infrastructure evolves to support humanity.