RECLAIMING THE MARKETPLACE OF IDEAS FROM THE DIGITAL CAULDRON OF ILLICIT LOVES

PROTECTING FREE SPEECH WHILE MODERATING CONTENT ON SOCIAL MEDIA PLATFORMS

Wendy K. Tam[1]

ABSTRACT

The Internet has become an indispensable part of modern life, facilitating, among other things, communication, work, news, and entertainment. The volume of user-generated content, particularly on social media, is mind-numbing. Harmful content within this vast collection of material proliferates, yet efforts to regulate online speech are stymied by First Amendment protections and Section 230 immunity. We propose a path forward with a time, place, and manner restriction on the volume of online speech—a proposal that aims to balance the benefits of online expression with the need to mitigate its harms by introducing a regulation framework that incorporates societal interests. Our approach offers a vision for a more sustainable digital ecosystem while promoting the foundational principles behind First Amendment free speech protections.

[1] Wendy K. Tam is Professor and Stevenson Chair, Departments of Political Science, Computer Science, and the Law School at Vanderbilt University and the Department of Biomedical Informatics at Vanderbilt University Medical Center. She is also an affiliate of the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

Table of Contents

INTRODUCTION

The Internet was "born" in the early 1990s.[2] Since then, it has morphed in simultaneously astonishing and unsettling ways, and so much so that one might easily imagine the Internet of today being described as science fiction just a couple of decades prior. We now reside in a hyperconnected online society that is heavily influenced by content that is continuously generated and viewed within our curated online networks. In 2023, Justice Thomas wrote, "[i]t appears that for *every minute* of the day, approximately 500 hours of video are uploaded to YouTube, 510,000 comments are posted on Facebook, and 347,000 tweets are sent on Twitter. On YouTube alone, users collectively watch more than 1 billion hours of video *every day*."[3] As well, "[i]n 2023, U.S. TikTok users uploaded more than 5.5 billion videos, which were in turn viewed more than 13 trillion times around the world."[4] Moreover, while this online content was largely human-generated a few decades ago, the content consumed today is generated by both humans and machines, further accelerating the rate, increasing the amount, and changing the character of the information that is produced and consumed.

Along with the explosive content has arisen a need and desire for content moderation on the social media platforms. Congress has attempted, and has had some success, in its efforts to pass legislation intended to reduce harmful online content. For example, in 2018, Congress passed FOSTA (Allow States and Victims to Fight Online Sex Trafficking Act),[5] legislation that made interactive computer services liable for content that promotes

[2] The "Internet" here refers to the World Wide Web, which began in 1993 with the introduction of Mosaic, the first graphical web browser. *See NCSA Mosaic (Project Highlights)*, Nat'l Ctr. for Supercomputing Applications, https://www.ncsa.illinois.edu/research/project-highlights/ncsa-mosaic/ (last visited Oct. 17, 2025). While the protocols that enabled the World Wide Web obviously existed prior to Mosaic, it was the graphical web interface that would make connectivity more accessible and widespread, opening the digital frontier to the broader public, and ultimately providing the spark that would ignite a technological revolution that would reshape the world.

[3] *Twitter, Inc. v. Taamneh*, 598 U.S. 471, 480 (2023).

[4] *TikTok, Inc. v. Garland*, 604 U.S. 56, 63 (2025).

[5] Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164, 132 Stat. 1253 (2018). The success of FOSTA in reducing the harms of sex trafficking has been challenged. Some

or facilitates sex trafficking, or the prostitution of another person. While politics has become more polarized in recent times,[6] congressional efforts in this arena, notably, have been characterized by bipartisan support and agreement that more must be done to moderate harmful content on social media platforms. These sentiments were on display in January 2024 at a Senate Judiciary Meeting with the CEOs of the large social media platforms where the focus was on social media content that has been harmful to children. Throughout the hearing, parents, who had lost their children to suicide, sat behind the CEOs, silently holding up pictures of their deceased children. In a particularly palpable and memorable moment, Josh Hawley (R–MO) pressed Mark Zuckerberg on whether he would like to apologize to the families sitting behind him. Zuckerberg turned to the families and said, "I'm sorry for everything you have all been through. No one should go through the things that your families have suffered."[7] The full committee hearing was chaired by Democrat Dick Durbin (D–IL) and Republican Lindsay Graham (R–SC). At the end of the hearing, Durbin declared, "I want to thank Senator Graham. He and I have differences on politics, but we sure do have a lot of things we agree on. This is one of them."[8]

Indeed, while Democrats and Republicans have disagreed about many issues in recent times, the need for content moderation on social media platforms has brought legislators on both sides of the aisle together to forge a path toward a solution. In the last couple of years, legislation that has been introduced with bipartisan co-sponsorship and intended to address online safety for children include:

contend that the situation for sex workers has been worsened, not ameliorated, by the legislation. Here, we are simply recognizing that there is agreement on a societal harm and some success in identifying a path that attempts to address a particular harm.

[6] *See*, e.g., Neil A. O'Brian, *The Roots of Polarization: From the Racial Realignment to the Culture* Wars (Univ. of Chicago Press, Chicago Press 2024), for a discussion of civil rights history and how it shaped partisan fault lines that continue to persist to today.

[7] *Recap: Senate Judiciary Committee Presses Big Tech CEOs on Failures to Protect Kids Online During Landmark Hearing*, U.S. Senate Comm. on the Judiciary (Feb. 2, 2024), https://www.judiciary.senate.gov/press/releases/recap-senate-judiciary-committee-presses-big-tech-ceos-on-failures-to-protect-kids-online-during-landmark-hearing.

[8] *Id.*

Proposed Legislation	Sponsor(s)	Republican Co-Sponsors	Democratic Co-Sponsors	Independent Co-Sponsors
Kids Online Safety Act[9]	Sen. Blumenthal (D–CT)	37	35	1
STOP CSAM Act[10]	Sen. Dick Durbin (D–IL)	3	2	
EARN IT Act[11]	Rep. Wagner (R–MO)	26	5	
	Sen. Graham (R–SC)	14	10	
SHIELD Act[12]	Sen. Klobuchar (D–MN)	4	4	
Project Safe Childhood Act[13]	Sen. Cornyn (R–TX)	9	10	
REPORT Act[14]	Sen. Blackburn (R–TN)	3	2	
Child Online Safety Modernization Act of 2023[15]	Rep. Ann Wagner (R–MO)	27	8	
DEFIANCE Act[16]	Sen. Dick Durbin(D–IL)	3	4	1
	Rep Alexandria Ocasio-Cortez (D–NY)	8	6	
COPPA[17]	Sen. Edward J. Markey (D–MA)	6	13	1
Sammy's Law[18]	Rep. Debbie Wasserman Schultz (D–FL)	10	8	

I. CONTENT MODERATION

Major platforms like X/Twitter (hereafter "Twitter" for simplicity), Facebook, Reddit, and Google are well aware that their platforms have hosted and continue to host harmful user-generated content.[19] And, although the platforms are granted broad-scale immunity for content that individual users produce and place on their platforms by Section 230 of the 1996 Communications and Decency Act (47 U.S.C. §230),[20] all of the major platforms, nonetheless, employ various mechanisms for content moderation. Why they moderate, when they have no legal obligation to do so, is generally believed to fall into three main

[9] Kids Online Safety Act, S. 1409, 118th Cong. (2023).

[10] STOP CSAM Act of 2023, S. 1199, 118th Cong. (2023).

[11] EARN IT Act of 2023, H.R. 2732, 118th Cong. (2023); S. 1207, 188th Cong. (2023).

[12] SHIELD Act of 2023, S. 412, 118th Cong. (2023).

[13] Project Safe Childhood Act, S. 1170, 118th Cong. (2023).

[14] REPORT Act, S. 474, 118th Cong. (2023).

[15] Child Online Safety Modernization Act of 2023, H.R. 5182, 118th Cong. (2023)

[16] DEFIANCE Act of 2024, S. 3696, 118th Cong. (2024); H.R. 7569 118th Cong. (2024). In addition to legislation aimed at the protection of children online, Congress has proposed a number of other laws aimed at reducing the scope of Section 230. For a discussion, *see* Meghan Anand et al., *All the Ways Congress Wants to Change Section 230: Republicans and Democrats Alike Want to Change Section 230 of the Communications Decency Act. Here's a Comprehensive List of the Proposed Legislation So Far*. SLATE (Mar. 23, 2021), https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html.

[17] Children and Teens' Online Privacy Protection Act, S.836, 119th Cong. (2025).

[18] Sammy's Law, H.R. 2657, 119th Cong. (2025)

[19] *See* Joel Kaplan, *More Speech, Fewer Mistakes*, META (Jan. 7, 2025). https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes, where Mark Zuckerberg discusses Meta's new content moderation strategy. He acknowledges that harmful content appears on the platform, but that this content falls within Meta's commitment to free expression.

[20] 47 U.S.C. § 230 (2018).

categories.[21] The first is a general commitment to American free speech norms. The second is a sense of corporate responsibility to positively impact society and adhere to ethical considerations, which may include advancing free speech norms while also reducing harmful content. The last, and not to be understated, is their own economic interests. For social media platforms, a central business goal is to create an engaging environment that incentivizes its users to spend significant time exploring the content on the platform. This includes creating a platform that complies with the norms that their users expect, which necessarily entails focused and extensive content moderation efforts.

In its early days (pre-2009), Facebook relied on a one-page list of internal "rules" for how content should be moderated.[22] Facebook's operational internal rules have evolved as the Internet has matured and content has proliferated. More recently, these documents became much longer, more nuanced, described multiple tiers of review, involved both automated procedures and manual review by humans, and included an escalation procedure that ends with a hearing by their Oversight Board.[23] This policy remains in flux. In January 2025, Zuckerberg announced sweeping changes that would coincide with his statement that "[w]e're going to get back to our roots and focus on reducing mistakes, simplifying our policies and restoring free expression on our platform."[24] Content moderation on Reddit follows more of a "self-regulating model" with platform-wide rules as well as community rules that are enforced by both subreddit moderators

[21] For a discussion, *see generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1456–68 (2011).

[22] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

[23] *See id.*; Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2232 (2020); Laurence R. Helfer & Molly K. Land, *The Meta Oversight Board's Human Rights Future*, 44 CARDOZO L. REV. 2233 (2023). *See also Facebook Community Standards*, META TRANSPARENCY CENTER., https://transparency.fb.com/policies/community-standards/(last visited Oct. 17, 2025).

[24] Joel Kaplan, *More Speech and Fewer Mistakes*, META (Jan. 7, 2025), https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes; Video posted by Mark Zuckerberg, META, *It's time to get back to our roots around free expression. We're replacing fact checkers with Community Notes, simplifying our policies and focusing on reducing mistakes. Looking forward to this next chapter*, (Jan. 7, 2025), https://www.facebook.com/watch/?v=1525382954801931.

and community members.[25] TikTok, as well, has its own community rules[26] and curated enforcement measures that employ algorithms as well as a human "trust & safety team."[27] Twitter once had a trust and safety team, but after Elon Musk purchased Twitter, he disbanded the team.[28] Twitter's content moderation policies are also in flux, but it has published a set of rules "to ensure all people can participate in the public conversation freely and safely" and allows users to set their own privacy settings, which provides a method by which users can customize their own content.[29] Twitter also utilizes a feature they call "Community Notes," intended to empower users "to collaboratively add context to potentially misleading posts" with "notes on any posts." These notes are shown publicly on the post if "enough contributors from different points of view rate that note as helpful."[30] Meta recently announced that they would be ending their third-party fact-checking program and adopting a similar "Community Notes model."[31] Meta maintains extensive documents in their Transparency Center[32] that list their policies that define what is and is not allowed on their platform,[33] their enforcement mechanisms,[34]

[25] *Reddit Rules*, REDDIT, https://redditinc.com/policies/reddit-rules (last visited Oct. 17, 2025).

[26] *Community Principles*, TIKTOK, https://www.tiktok.com/community-guidelines/en/community-principles (last visited Oct. 17, 2025).

[27] *Enforcement Guidelines*, TIKTOK, https://www.tiktok.com/community-guidelines/en (last visited Oct. 17, 2025).

[28] Twitter Trust and Safety Council, *Joint Statement on the Disbanding of the Twitter Trust and Safety Council*, CTR. FOR DEMOC. & TECH. (Dec. 14, 2022); Patience Haggin, *Elon Musk's Twitter Disbands Trust and Safety Council,* WALL ST. J. (Dec. 12, 2022), https://www.wsj.com/articles/elon-musks-twitter-disbands-trust-and-safety-council-11670898329.

[29] *The X Rules,* X, https://help.x.com/en/rules-and-policies/x-rules (last visited Oct. 17, 2025); *Privacy Policy,* X, https://privacy.x.com/en (last visited Oct. 17, 2025).

[30] *About Community Notes on X,* X, https://help.x.com/en/using-x/community-notes (last visited Oct. 17, 2025).

[31] Joel Kaplan, *More Speech, Fewer Mistakes*, META (Jan. 7, 2025), https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

[32] META TRANSPARENCY CENTER, https://transparency.meta.com/(last visited Oct. 17, 2025).

[33] *Policies*, META TRANSPARENCY CENTER, https://transparency.meta.com/policies/ (last visited Oct. 17, 2025).

[34] *How Meta Enforces Its Policies*, META TRANSPARENCY CENTER, https://transparency.meta.com/enforcement/ (last visited Oct. 17, 2025).

their community standards,[35] and their governance.[36] There is no single approach or framework for content moderation on the various social media platforms. Instead, each platform has evolved its own strategies that are customized for its own business objectives and its particular user base. To be sure, even without legislation that would provide a legal incentive, the platforms have salient economic incentives to provide a favorable user experience, which, in part, requires a significant level of effort directed toward content moderation.

However, despite concrete and concentrated efforts by social media platforms to moderate their content, the evolving nature as well as the sheer scale and speed of the continuously and rapidly generated user content present non-trivial difficulties for employing responsive, effective, and scalable moderation strategies. Indeed, there remains a general unhappiness and unease by users and legislators alike who continue to easily identify copious harmful content that appears to deftly evade the various, and even extensive, content moderation procedures that are currently employed by social media platforms.

A. Impediments to Removing Harmful Online Content

Without making value judgments about whether any particular piece of content on the Internet is harmful or not, we begin from an uncontroversial premise—despite efforts by Congress as well as the social media platforms, *some* of the content that continues to find its way onto social media platforms is harmful to society. We have discussed congressional efforts aimed at the clearly important goal and imminent need of safeguarding our children from exploitation. However, the societal harms emanating from online platforms, sadly, are more extensive.

[35] *See Community Standards*, META TRANSPARENCY CENTER, https://transparency.meta.com/policies/community-standards/ (last visited Oct. 18, 2025).
[36] *See Governance*, META TRANSPARENCY CENTER, https://transparency.meta.com/governance/ (last visited Oct. 18, 2025).

Another area where there seems to be general agreement about societal harm stems from manifestations of coordinated inauthentic behavior (CIB), a term coined by Facebook, to refer to "coordinated campaigns that seek to manipulate public debate across our apps." [37] CIB is an umbrella term that encompasses any tactic by which inauthentic behavior is coordinated. By their own telling, Facebook expends significant resources toward actively monitoring and removing content it regards as CIB. These types of events are sufficiently frequent that Facebook issues a quarterly report of their activity and efforts in thwarting these types of adversarial threats that manifest on their platform.[38] Facebook has identified Internet bots (which are generally understood as software written by a human to automate tasks on the Internet) that were at the forefront of an effort by Russian propaganda efforts to amplify content favorable to Donald Trump prior to the 2016 election.[39] In addition to this high profile CIB incident, Facebook reports taking down other CIB networks linked to governments or ruling parties in Serbia, Cuba, and Bolivia, cyber espionage from South Asia, and covert influence operations in Venezuela, Iran, China, Georgia, Burkina Faso, and Togo, to name a few.[40] They found these adversarial threats from other countries to be malicious in nature and intended toward instigating societal harm in the US.

[37] *See Coordinated Inauthentic Behavior*, META, https://about.fb.com/news/tag/coordinated-inauthentic-behavior/ (last visited Oct. 18, 2025).

[38] Their quarterly adversarial threats report (*e.g.*, Ben Nimmo, *Meta's Adversarial Threat Report, Fourth Quarter 2022*, META (Feb. 23, 2023), https://about.fb.com/news/2023/02/metas-adversarial-threat-report-q4-2022/) regularly reports on the outcome of their efforts to control CIB. Meta aptly describes this as an "ongoing effort" where they are "committed to continually improving to stay ahead."

[39] *See Update on Twitter's review of the 2016 US election*, X BLOG (Jan. 19, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html; Gerrit De Vynck & Selina Wang, *Russian Bots Retweeted Trump's Twitter 470,000 Times*, BLOOMBERG (Jan. 26, 2018), https://www.bloomberg.com/news/articles/2018-01-26/twitter-says-russian-linked-bots-retweeted-trump-470-000-times.

[40] *See* Ben Nimmo, *Meta's Adversarial Threat Report, Fourth Quarter 2022*, META (Feb. 23, 2023), https://about.fb.com/news/2023/02/metas-adversarial-threat-report-q4-2022/. *See also Building the Future of X*, X BLOG (Aug. 2, 2023), https://blog.x.com/, for a similar page maintained by Twitter that describes various content moderation philosophy and efforts.

i. Legal Hurdles for Moderating Harmful Content

While it is uncontroversial that some social media content is harmful, how we might mitigate the unrelenting multi-directional and multi-faceted onslaught of these malicious efforts is anything but straightforward. Laws in the US pose substantial legal constraints for possible constitutional paths forward. Consider that *if* we were to make a law regulating social media content, that law would likely implicate free speech, and so any such law must be carefully crafted to pass constitutional muster. Indeed, regulation in the social media space requires a scalpel, not a bludgeon, to constitutionally navigate around free speech protections.

Any law that implicates free speech almost surely needs to be content-neutral, which means that the approach cannot be based on value-laden judgments about whether certain content or a particular coordination effort is "good" or "bad." If one is specifically targeting CIB, choosing which coordinated speech to moderate in a content-neutral way can be challenging. Generally, users coordinate to amplify the reach and visibility of specific content. Coordination might take, in its simplest form, organizing retweet efforts on Twitter or likes on Facebook posts. Of course, coordination, in and of itself, does not imply that the content is somehow harmful. The coordination may simply be raising the visibility of non-harmful content such as a picture of Barack and Michelle Obama on their anniversary or a charity fundraising campaign. Distinguishing simple crowd behavior from CIB is difficult, in part, because the prima facie behavioral patterns are similar, with only the underlying and unknown motivations that set them apart.

Certainly, one way to distinguish simple coordination from CIB is via value-laden judgment calls. The issue with using value-laden judgments for determining the motivation behind coordination efforts, however, is obvious—these judgments may be incorrect. Moreover, and critically, these judgments are not content-neutral. They are content-based and value-laden, which does not portend objectivity. To muddy the waters further, even coordination *with* inauthenticity does not necessarily imply that the coordination effort is for harmful content. For instance, fake reviews and fabricated testimonials are a common tactic for businesses seeking to boost engagement for their advertisements.[41] While such practices may be distasteful or undesirable, the content itself is not necessarily harmful.[42] To make matters even more difficult, Douek identifies another critical problem that stems from the fact that "[c]oordination and authenticity are not binary states but matters of degree, and this ambiguity will be exploited by actors of all stripes."[43] Adding to the complexity, Facebook's definition of CIB is constantly evolving,[44] and, to some extent, depends on proprietary information about how it implements its recommendation algorithms. In short, identifying harmful CIB content in a content-neutral way appears to be all but futile.

Besides being content-neutral, a law implicating free speech needs to be narrowly tailored to serve a compelling governmental interest, such as, for example, national security or public safety. Designing a narrowly tailored approach to address these governmental interests without overreaching or underreaching is a complex and difficult task. That is, the method must fairly accurately identify harmful content without

[41] *See* Press Release, FED. TRADE COMM'N, *Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials* (Aug. 14, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials.

[42] Buying engagement on the online platforms is not a new phenomenon. *See, e.g.*, Charles Arthur, *How low-paid workers at 'click farms' create appearance of online popularity*, THE GUARDIAN (Aug. 2, 2013), https://www.theguardian.com/technology/2013/aug/02/click-farms-appearance-online-popularity (describing "click farms"). This practice is actively monitored by the platforms by various auditing techniques. *See, e.g.,* Dave Lee, *Instagram deletes millions of accounts in spam purge*, BBC NEWS (Dec. 19, 2014), https://www.bbc.com/news/technology-30548463; Jim Edwards, *Facebook Targets 76 Million Fake Users in War On Bogus Accounts*, BUS. INSIDER (Mar. 5, 2013), https://www.businessinsider.com/facebook-targets-76-million-fake-users-in-war-on-bogus-accounts-2013-2. This war has intensified with time. Meta now reports deleting billions of fake accounts each year. *See Fake Accounts*, META TRANSPARENCY CENTER, https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/ (last visited Oct. 22, 2025).

[43] Evelyn Douek, *What Does "Coordinated Inauthentic Behavior" Actually Mean?*, SLATE (Jul. 2, 2020), https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html.

[44] *See Inauthentic Behavior*, META TRANSPARENCY CENTER, https://transparency.meta.com/policies/community-standards/inauthentic-behavior/ (last visited Oct. 25, 2025). *See also* Nathaniel Gleicher, comment on X post (June 21, 2020, at 2:54 PT), https://twitter.com/ngleicher/status/1274823045031460864 (discussing whether a prank to inflate tickets to a Trump rally constituted CIB).

mistakenly including content that should be excluded or excluding content that should be included. Herein lies not only a legal hurdle, but a technical hurdle. Once one accomplishes this monumentally difficult task of defining which online content is harmful without using value-laden judgments, how would one then design a technological solution that is able to reliably distinguish that harmful content from not harmful content in a sufficiently precise way to pass constitutional scrutiny? Indeed, we have, not one, but two seemingly impossible tasks that must be achieved within our strong constitutional constraints.

ii. Technical Hurdles for Moderating Harmful Content

Indeed, regulation in the content moderation space faces substantial legal hurdles which are intertwined with equally thorny technical quagmires. Consider that *if* we were able to come to an agreement that particular online content is harmful, developing scalable and effective methods for identifying such content presents its own non-trivial technical task. To be sure, distinguishing "good bots" from "bad bots," or differentiating simple crowd behavior from harmful CIB, which is also crowd behavior, but of an ilk that tends toward amplifying propaganda or misleading information, are not simple tasks even for humans to agree upon, much less to then develop a system to infuse these nuanced human values into some type of machine automation.[45] While humans may have been able to moderate content in the very early days of social media, the need for

[45] Meta reports that they remove content that engages in 1) mass reporting, i.e., "adversarial networks that coordinate to abuse our reporting systems to get accounts or content incorrectly taken down from our platform, typically with the intention of silencing others", 2) brigading, i.e., "adversarial networks that work together to engage in repetitive behavior, often in the form of sending direct messages to their targets, or mass-commenting on their posts", and 3) coordinated violating networks, i.e., "coordinated violating networks when we find people—whether they use authentic or fake accounts—working together to violate or evade our Community Standards." The factors they consider in these determinations include "coordination signals", "high volume of reports", "misleading & abusive nature of reports (e.g. reporting innocuous posts as violating)", "repetitive targeting to harass or silence people, usually with unsolicited messages or comments," "systematic violations of our Community Standards", and "efforts to evade enforcement."

automation is now a foregone conclusion given that "Facebook users, for example, share more than 100 billion messages every day."[46]

To complicate matters, we have a moving target. Although Facebook already actively monitors and tries to reduce harmful CIB content, Facebook purposefully does not share "the exact thresholds and precise signals we rely on to tackle this abuse" "to avoid tipping off these groups."[47] Indeed, a fairly precise definition of this harmful content is, itself, its own death knell because it would provide bad actors with the exact parameters to circumvent, creating a whack-a-mole scenario that further exacerbates the challenge of identifying a technical solution.[48] Already, Meta purports to expend significant resources toward content moderation. In 2024, they reported spending $5 billion on safety and security, with 40,000 people working on it.[49] Nevertheless, despite these substantial efforts, we know undesirable behavior continues to seep through the Meta sieve. Indeed, the construction of any such technical filter is fraught with difficulty.

There exist some content classifications where we have been able to achieve broad-scale agreement on harm. For instance, child sexual abuse material and other types of child exploitation fall into this category. But even in these areas of general agreement, devising a technological tool that is effective at scale remains a daunting task.[50] Simply sifting and parsing through the enormous magnitude of content on the Internet to reliably identify *any* category of content is a formidable task that necessitates significant and

[46] *See Moody v. NetChoice, LLC* and *NetChoice, LLC v. Paxton*, 603 U.S. 707 (2024).

[47] Ben Nimmo et al., QUARTERLY ADVERSARIAL THREAT REPORT: FOURTH QUARTER 2022, at 11 (Feb. 2023), https://about.fb.com/wp-content/uploads/2023/02/Meta-Quarterly-Adversarial-Threat-Report-Q4-2022.pdf.

[48] *See, e.g.*, Christopher Knaus et al., *Inside the hate factory: how Facebook fuels far-right profit*, THE GUARDIAN (Dec. 5, 2019), https://www.theguardian.com/australia-news/2019/dec/06/inside-the-hate-factory-how-facebook-fuels-far-right-profit.

[49] *See* Meta, *Our Work to Help Provide Young People with Safe, Positive Experiences* META (Jan. 31, 2024), https://about.fb.com/news/2024/01/our-work-to-help-provide-young-people-with-safe-positive-experiences/.

[50] There is some progress in these areas. For instance, PhotoDNA (*PhotoDNA*, MICROSOFT, https://www.microsoft.com/en-us/photodna (last visited Oct. 22, 2025)) is a technological tool developed to identify images of child exploitation. This tool is widely used to combat child exploitation online. It creates unique hashes from a database of known images and video files and compares those to other images.

constant technical innovation. To be sure, protecting free speech is a fundamental and challenging hurdle that must be surmounted, but even after we identify a viable legal path forward, we must still craft a scalable and effective technological solution that aligns with constitutional requirements.

III. Viable Legal Paths Forward

Laws that restrict speech based on its content are subject to strict scrutiny. While some laws pass strict scrutiny, the strict scrutiny standard is extremely rigorous and very often fatal. We propose to bypass the rough terrain that is forewarned by a strict scrutiny analysis by traversing the slightly less daunting path that is paved for content-neutral speech restrictions. In particular, we propose that a viable path forward exists in a **time, place, and manner restriction** on speech.[51] In *Ward v. Rock against Racism*, the Supreme Court ruled that while music is an expressive and protected form of speech, the volume of amplified music in Central Park could be regulated without violating free speech rights. In this case, the city of New York had adopted a regulation that required performers at an amphitheater in Central Park to use sound amplification equipment and a sound technician provided by the city. This regulatory measure was passed after the city had received numerous complaints of excessive noise. The Court found that the regulation on sound amplification was a time, place, and manner restriction on speech, and not intended to control the content of the speech. Justice Kennedy, writing for the majority, clarified that a time, place, and manner regulation "must be narrowly tailored

[51] We recognize that time, place, and manner restrictions can be used to regulate speech on public property, but social media is run by private companies. As such, the institutional setting of social media is not a traditional public forum. *See*, however, Elon Musk (@elonmusk), "Twitter serves as the de facto public town square", X (Mar. 26, 2022, at 10:51 PT), https://x.com/elonmusk/status/1507777261654605828); *Packingham v. North Carolina*, 582 U.S. 98, 107 (2017) (where the Court wrote that social media platforms have become the "modern public square."); Mary Anne Franks, *Beyond the Public Square: Imagining Digital Democracy*, 131 YALE L.J. 427 (2021); 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). However, because the institutional setting is increasingly publicly purposed though privately owned and monetized, the Court may consider formalizing its institutional setting to be a hybrid setting that is not fully public but not fully private.

to serve the government's legitimate, content-neutral interests but that it need not be the least restrictive or least intrusive means of doing so."[52] In short, the case established that time, place, and manner restrictions must satisfy a four-pronged test. First, they must be content-neutral. Second, they must serve an important governmental interest. Third, they must be narrowly tailored. Fourth, they must provide ample alternative means for communication.[53]

Along these lines, we propose a focus on the *volume of social media speech* rather than on the particular content of social media speech.

Considering time, place, and manner restrictions seem particularly apt when we consider that one of the things that social media has fundamentally changed is the *manner* in which speech now manifests, which *then* shapes the content. Viewing online speech in this way highlights the need for structural interventions that address the root causes of online discourse dysfunction. Since a time, place, and manner restriction that targets the volume of online speech rather than the content of that speech is **content-neutral**, we are able to bypass any need to make value-laden judgments about whether the content of the speech is "good" or "bad." Note that speech on social media platforms has two possible "speakers" or entities that affect how online speech manifests. The content can be seen as the speech of the person who posted the content. While one might conceive of the original content as the speech of the user, the "volume" of that speech is created by the platform. Everyone is able to post content, but some content is amplified by the platform while other content is not amplified. In this sense, since the volume is controlled

[52] *Ward v. Rock Against Racism*, 491 U.S. 781, 798 (1989); *Kovacs v. Cooper* 336 U.S. 77 (1949) (upholding an ordinance prohibiting the use of sound trucks that emitted "loud and raucous"" noises on city streets); *McCullen v. Coakley*, 573 U.S. 464, 471 (2014) (where the Court considered a law barring individuals from entering or remaining "on a public way or sidewalk adjacent to a reproductive health care facility within a radius of 35 feet." Justice Roberts wrote that "[t] he buffer zones burden substantially more speech than necessary to achieve [Massachusetts'] asserted interests.")

[53] *Ward*, 491 U.S. at 802.

and determined by the platform, that aspect of the speech could be considered as originating from the platform.

We note as well that, in *Moody*, the Court stated that "When the platforms use their Standards and Guidelines to decide which third-party content those feeds will display, or how the display will be ordered and organized, they are making expressive choices. And because that is true, they receive First Amendment protection."[54] So, if the manner in which user content is displayed is considered the expressive product of a platform, then that editorial curation by the platform, as the Court stated in *Moody*, is protected by the First Amendment. However, protected speech may still be subject to time, place, and manner restrictions.[55]

Time, place, and manner restrictions on speech are designed to strike a balance between competing societal values. So, while freedom of speech holds a central place in our society, so too do **important governmental interests** like social order, civil discourse, and public safety. The Court also values a vibrant exchange of ideas and writes that "[s]tates (and their citizens) are of course right to want an expressive realm in which the public has access to a wide range of views. That is, indeed, a fundamental aim of the First Amendment."[56] However, the Court cautions that, in pursuit of these First Amendment goals, the First Amendment also prevents the government from "tilt[ing] public debate in a preferred direction"[57] In addition, the government cannot prevent private actors, which may include social media platforms and its users, "from speaking as they wish and preferring some views over others. And that is so even when those actors possess 'enviable vehicle[s]' for expression."[58] At the same time, the Court acknowledges that

[54] *Moody*, 603 U.S. at 740.
[55] Free speech issues surrounding algorithms is a complex legal arena without, as of yet, firm direction from the Supreme Court. However, whether or not these algorithms will or will not be protected speech, we know that even protected speech can be regulated with time, place, and manner restrictions.
[56] *Moody*, 603 U.S. at 711.
[57] *Sorrell v. IMS Health Inc.*, 564 U.S. 552, 556 (2011).
[58] *See Moody*, 603 U.S. 707, 733 (2024) (citing *Hurley v. Irish-American Gay, Lesbian and Bisexual Grp. of Bos.*, 515 U.S. 557, 577 (1995)).

while restrictions whose *purpose* is to "rejigger the expressive realm"[59] are likely to be unconstitutional, the Court is sympathetic that "[i]n a better world, there would be fewer inequities in speech opportunities; and the government can take many steps to bring that world closer."[60]

These declarations by the Court underscore its commitment not only to the textual protections of the First Amendment but also to the underlying *values* those protections are meant to secure—individual autonomy and democratic self-governance. Consistent with this commitment, the Court has long recognized that when the government operates within the narrow boundaries permitted by the First Amendment, it may adopt measures that help sustain a speech environment in which those foundational values can flourish. Time, place, and manner restrictions, properly tailored, reflect this approach. They do not suppress content or ideas, but, instead, structure the conditions of expression in ways that strengthen, rather than weaken, the principles that animate robust free-speech rights.

In the absence of a mechanism to manage online speech volume, we risk allowing that volume to distort the very values the First Amendment aims to protect. When a small number of individuals dominate online platforms, this imbalance can undermine free speech principles by impairing the autonomy of listeners—an essential component of free speech theory—as audiences are compelled to listen to conversations that are disproportionately shaped by a select few.[61] As well, it implicates and impairs the right of expressive association or the "right to associate" with one's preferred speakers.[62] While one might still be able to find particular posts by particular speakers, the flood of platform-curated content inevitably shapes one's expressive communities and hinders

[59] *Moody*, 603 U.S. at 733.

[60] *Id.* at 741.

[61] *See Martin v. City of Struthers*, 319 U.S. 141 (1943) (on the right to listen); *Stanley v. Georgia*, 394 U.S. 557 (1969); James Grimmelmann, *Listeners' Choices Online*, SO. CAL. L. REV. 2025 (discussing listener choices online and how this perspective transforms how we might understand First Amendment protections for social media).

[62] *See Rumsfeld v. Forum for Acad. and Institutional Rts., Inc.*, 547 U.S. 47 (2006) (on the right of expressive association).

this right of expressive association. Importantly, a time, place, and manner approach aims not to silence any individual voice, to change the ideological balance of the debate, or to prevent platforms from continuing to amplify or promote the speech of various individuals. Rather, it aims to manage online speech volume so that the online speech environment more faithfully reflects and advances the First Amendment's core commitments by promoting the conditions that free speech protections are designed to secure.

What might be the offline equivalent of our online scenario where some content is amplified over other content?

> Imagine a large public forum, like a vast park, touted as a "free speech haven." Anyone can enter, wander around, and speak freely. Parks have always existed but never at this scale where people are able to gather, chat, debate, and share their views without restriction. The park never closes and offers something for every interest, no matter how obscure. The park is much like other parks, except, there is an interesting twist: an unseen entity, the Park Overlord, is always observing, and subtly shaping the experience of every visitor.
>
> When you first step into the park, the Park Overlord astutely notes your interests. If you express enthusiasm for a particular topic, the Park Overlord immediately directs you toward a group of like-minded individuals. You are thrilled as you join the group and find it to be a welcoming and vibrant environment, teeming with energy. You thought you knew everything about the topic, but hearing more perspectives and being exposed to new ideas from so many interesting people fascinates you and piques your curiosity.
>
> In this group, everyone is free to speak, but the Overlord controls the volume of every speaker. Some people are given powerful megaphones so that their voices are booming across the group, while others are restricted to speak only in low tones that are nearly inaudible. Even though everyone can technically "speak freely," the loudest voices quickly come to dominate what you hear. And these voices? They tend to echo the ideas you already agree with, reinforcing your views. Despite finding it odd that you have lost control over the volume of your voice, you feel happy, seen, and comfortable.

You are, of course, free to wander about the park and explore other groups. The park is not restrictive at all in this sense. There are no walls between groups of people, so you are free to wander over to any gathering and join in. However, if you linger with any group for even a short while, the Park Overlord notices and begins nudging you to stay. You are quickly welcomed to any group you choose, and the Overlord reinforces and validates your choices by making it easy to find similar voices. You are so enthralled that you do not even notice that, as time wears on, other groups become slightly but surely less prominent.

It is not until you spend considerable time reflecting that you realize that the park is strangely dystopian. At first, it seemed to be the true realization of a free speech haven where everyone has the same opportunity to speak, listen, and explore. Technically, this is true—everyone is treated identically in this respect. No one is ever silenced; all are free to speak; and all are guided toward like-minded others. The Overlord genuinely wishes for all park visitors to feel happy, comfortable, and engaged. But the Overlord's total and absolute control over how each person's speech is amplified has a profound effect. By deciding who receives megaphones and who must whisper, he subtly but powerfully shapes your mind and thoughts. The amplified voices dictate group discussions and influence your understanding of the topic. Other voices, though inarguably present, are effectively unheard. The right to listen—to hear a variety of perspectives—is subtly but surely impaired. The appearance of open dialog conceals the reality of a distorted soundscape of manipulated conversations.

Perhaps "free speech" is not just when everyone is free to speak, without consideration of the conditions that facilitate the free and unfettered exchange of ideas. Instead, "free speech" includes not only the freedom to express oneself but also the mechanisms by which speech is heard, understood, and engaged with. It involves considering how ideas are circulated, who has access to them, and the dynamics of interaction with platforms, audiences, and institutions that are structured in ways that promote meaningful dialog. A meaningful exchange of ideas requires mechanisms that allow all participants to have the ability to engage effectively. Without such conditions, the right to listen and the right to expressive association is compromised. One is, instead, compelled to listen to and associate with a select few.

It is perhaps not subtle that this fictitious park mimics the way online platforms use algorithms to recommend content, spotlight certain voices, and guide conversations. Like the Park Overlord, these algorithms do not censor outright, allow all to speak, but yet amplify, quietly deciding who

> gets a megaphone and who must whisper. In doing so, the platforms shape the conversation, not by limiting speech, but by controlling what we hear most loudly.

Surely, all harmful online speech is not attributable to volume, and we do not purport to be identifying all forms of harmful online speech. There is no requirement that we eradicate all evils of the same genus or none at all.[63] The magnitude of the problem we are addressing is daunting. Fully addressing it will require ingenuity and a multifaceted approach. Here, we are offering just one potential path, but also arguing that focusing on the speech volume that is created by social media recommendation algorithms is a viable legal route that would be particularly helpful in reducing the harms produced by speech on the social media platforms. A crucial observation in this regard is that a distinct feature of online speech is that the potential harm of online speech is commensurate with its volume. Without volume, the potential harms from *any* type of online content are mitigated. For instance, CIB is not effective unless the volume of its resulting speech is high; and as the volume increases, so too does the potential harm to society, regardless of content. Indeed, the point of coordination is to increase the volume of the speech. Low levels of coordinated speech lack the capacity to cause the same level of harm that higher levels of coordination are capable of producing. Quite clearly, the degree of harm and the effectiveness of tactics like CIB, bots, and botnets[64] are positively correlated with increasing volume. Indeed, *it is precisely the volume that is the essence of the harm*.[65]

[63] *See Ry. Express Agency v. New York*, 336 U.S. 106 (1949) (discussing rational scrutiny standard).

[64] Botnet is a portmanteau, formed from a combination of the words “robot” and “network.” It refers to a network of devices that are centrally controlled, usually for nefarious purposes.

[65] *See also Barr v. Am. Ass'n of Pol. Consultants, Inc.*, 591 U.S. 610, 636 (2020) (allowing a total content-neutral ban on robocalls in an analysis that indicated that it was the potential volume of such calls that was irksome and therefore could be regulated).

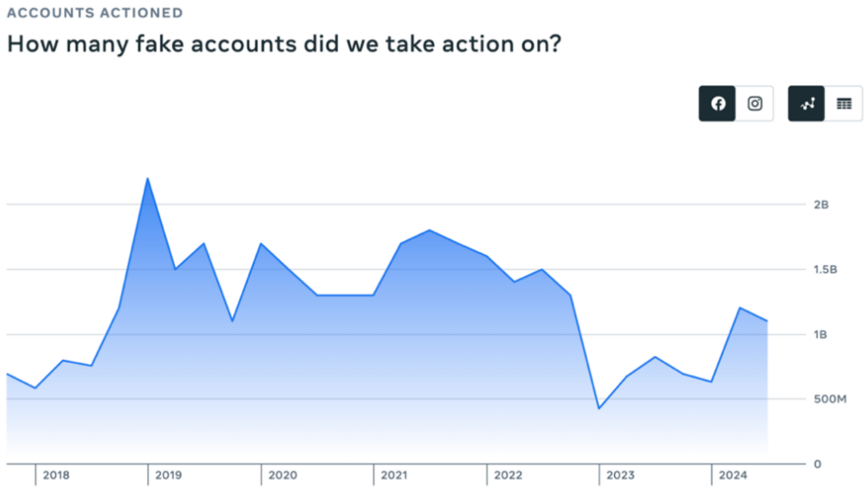

Figure 1. **Fake Accounts removed on Facebook every quarter. Figure reproduced from Meta Transparency Center's report on Fake Accounts. See https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/**

To boot, and without dispute, malicious actors have identified this vulnerability and are actively exploiting it, again highlighting how speech volume (rather than content) is the intrinsic harm. To be sure, malicious bots and botnets are innovations designed specifically to exploit the way speech unfolds on social media platforms. They would not exist *but for* the design of social media algorithms. Large-scale bot networks played a role in Russia's interference in the 2016 election, where coordinated social media accounts followed, shared, and retweeted each other to skew discourse and stir up discord by artificially inflating and spreading specific narratives.[66] This type of information propagation is initialized within node clusters on the social network, with the goal of injecting and propagating divisive narratives. The botnets manufacture artificial interactions and create the illusion of large-scale "engagement," which encourages platforms to further amplify and enable the reach of that content in a perpetual and self-reinforcing feedback loop.

[66] *See* Taylor Hatmaker, *Special Counsel Robert Mueller Indicts Russian Bot Farm for Election Meddling*, TECHCRUNCH (Feb. 16, 2018), https://techcrunch.com/2018/02/16/mueller-indictment-internet-research-agency-russia/.

It is unclear how significant the problems caused by fake accounts on social media truly are, but the number of fake accounts is staggering, and particularly striking when compared to the number of active monthly users. On Facebook alone, in the first three months of 2019, Meta removed 2.2 *billion* fake accounts.[67] As we can see from Figure 1 (reproduced from Meta Transparency Center's report on Fake Accounts), they regularly remove billions of fake accounts every year.[68] Compare this with the estimate of the total number of active monthly users on Facebook—approximately 3 billion at the end of 2023.[69] The onslaught and amount of speech from these fake accounts is staggering. Although the precise number is unknown, Twitter likewise has a substantial number of bots on their platform, up to 15% of all Twitter users, by some accounts.[70] Embedding a botnet infrastructure that one controls within the larger network connectivity structure enables manipulation of the information flow on social platforms by capturing control of the social media recommendations algorithms that boost content visibility.[71] This not the free exchange of ideas envisioned by our Founding Fathers, but rather, an attack on our ability to have a free exchange of ideas amidst this technological war. Managing online

[67] *See* Fake Accounts, META TRANSPARENCY CENTER https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/ (last visited Oct. 23, 2025).

[68] *Id.*

[69] *See* Robin Geuens, *How many users does Facebook have?*, SOAX (Oct. 30, 2024), https://soax.com/research/how-many-users-does-facebook-have; Simon Kemp, *Facebook Users, Stats, Data & Trends*, DATAREPORTAL (Mar. 12, 2023), https://datareportal.com/essential-facebook-stats; *Most popular social networks worldwide as of April 2024, by number of monthly active users*, STATISTA (Feb. 2025), https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[70] Jorge Rodríguez-Ruiz et al., *A One-Class Classification Approach for Bot Detection on Twitter*, 91 COMPUT. & SEC. 101715 (2020).

[71] *See* Lutz Finger, *Do Evil—The Business of Social Media Bots*, FORBES (Feb. 17, 2015), https://www.forbes.com/sites/lutzfinger/2015/02/17/do-evil-the-business-of-social-media-bots/; SAMUEL WOOLLEY, MANUFACTURING CONSENSUS: UNDERSTANDING PROPAGANDA IN THE ERA OF AUTOMATION AND ANONYMITY, (Yale Univ. Press 2023); NICK MONACO & SAMUEL WOOLEY, BOTS (Polity Press 2022). Of course, this is not the only path toward harmful speech volume. More recently, federal prosecutors and the Department of Justice identified new tactics. Instead of utilizing bot networks and fake accounts, they have accused two Russians of attempting to influence American public opinion and advance Russia's geopolitical goals by exploiting American social media influencers. *See* Steven Lee Myers, Ken Bensinger, & Jim Rutenberg, *Russia Secretly Worms Its Way Into America's Conservative Media*, NEW YORK TIMES (Sept. 7, 2024), https://www.nytimes.com/2024/09/07/business/media/russia-tenet-media-tim-pool.html.

speech volume contributes to public safety by helping to dispel the harm from these types of societal attacks.

Our offline scenario with the Park Overlord highlights how the park dynamics are ripe for exploitation by bad actors, and how the design of social media algorithms focused on engagement embodies a vulnerability that compromises both free speech ideals as well as public safety.

> Far away from the park, in a shadowy boardroom, a group of nefarious operatives gathers around a dimly lit table. They also have been fascinated by the popularity of the park, and so begin observing the park from afar, studying the patterns and behaviors of the visitors, as well as how the Park Overlord manages the park. They discover a vulnerability in the park's ecosystem—a flaw in the Park Overlord's pursuit of engagement above all else. The Park Overlord's management enables foreign manipulation and exploitation.
>
> The operatives prepare an ambitious plan. They begin by developing robots—machines indistinguishable from human park visitors. These robots laugh at the right moments, comment on videos and photos, and wander through the park just like any other human visitor. Millions of robots are created with the goal of gaining control of the megaphones and thus seizing control of the park conversations.
>
> The robots engage strategically. They create divisive content and share it, at first, only with other robots. They know that the Park Overlord will not give them a megaphone unless they engage the other park visitors. They learn that saying divisive and sensational things garners attention and enables them to steer conversations toward perspectives and narratives of their choosing. The Park Overlord, eager to keep visitors engaged, unwittingly, but willingly, hands megaphones over to the robots.
>
> Once the robots establish control, the operatives' plans begin to unfold. They use the megaphones to stoke divisiveness among the park visitors. Their speech is carefully crafted and designed to provoke outrage and sow discord. The impact is profound. Visitors, once so comfortable with their like-minded peers, now discover that there are other park visitors who have profoundly different and offensive views. The once joyful atmosphere turns tense and combative. The Park Overlord, blind to the true source of the chaos, continues to optimize for engagement, inadvertently but continuing to enable and even fuel the operatives' agenda.

> The operatives watch with satisfaction. The park, now a tool for manipulation, become a means to influence public opinion and destabilize the fabric of society. They introduce narratives that erode trust in institutions, pit communities against each other, and distract visitors from meaningful issues. The park, once a symbol of free speech, is now the battleground for a covert information war.
>
> Amidst the chaos, a few park visitors begin to notice the shift. They sense that something isn't right. Uncovering why is no easy task. The undercover robots are now everywhere, their presence hidden in plain sight, their influence woven into the very fabric of the park. As the park descends further into turmoil, the question becomes: can the visitors reclaim their spaces from those who seek to control them? Or will the park remain a battleground, its potential for good overshadowed by the forces of manipulation and division?

Indeed, not only have bad actors *already* infiltrated vulnerabilities on social media platforms, but their tactics also continually adapt and become increasingly difficult to detect. At a Senate hearing titled "Terrorism and Social Media: #IsBigTechDoingEnough?," Sen. John Thune (R–SD) began the meeting stating the positive contributions of social media, but then noted because of our free speech commitments and the light regulatory policy for Internet service providers, the "enemies of our way of life have sought to take advantage of our freedoms to advance hateful causes. Violent Islamic terrorist groups like ISIS have been particularly aggressive in seeking to radicalize and recruit over the Internet and various social media platforms."[72] Focusing on the volume of online speech that is controlled by social media algorithms thus serves multiple important governmental interests. On the public safety front, it helps us safeguard national security by curbing the ability of foreign adversaries who seek to dominate our information ecosystem, manipulate our discourse, and sow society

[72] *Terrorism and Social Media: #IsBigTechDoingEnough?,* Hearing Before the S. Comm. on Commerce, Sci., & Transp., 115th Cong. (Jan. 17, 2018), https://www.commerce.senate.gov/2018/1/terrorism-and-social-media-isbigtechdoingenough. Consider also that this content allows surveillance efforts by U.S. intelligence agencies, so the best way to moderate this content is not as seemingly straightforward as it might seem at first blush.

discord. We can see from Meta's quarterly adversarial threat reports that these threats are massive and relentless.[73]

Apart from public safety, managing online speech volume also restores individual autonomy to choose what they wishes to listen to, as they are not strongly confined to listening to content that is dictated by platform algorithms but can more easily access both more diverse and preferred viewpoints. Moreover, it supports the freedom of expressive association by reducing the non-subtle pressures that nudge users toward engaging with the most amplified voices, which distorts authentic connections and inhibits the ability of individuals to choose how to unite to collectively express, promote, pursue, or defend their shared ideas, beliefs, or interests. Ultimately, managing online speech volume helps to cultivate the robust exchange of ideas that serves democratic self-governance and sits at the helm of a healthy democracy.

Recommendation algorithms are fundamentally designed to maximize user engagement by serving each individual content that aligns closely with their preferences. Even when the content itself is benign, say, wellness or cat videos, this engagement-maximizing design can be harmful. By reinforcing emotionally stimulating material, these algorithms can foster addictive patterns of use, as users become locked into dopamine-driven feedback loops.[74] Empirical research links such sustained engagement to adverse mental health outcomes, including anxiety, depression, and diminished self-esteem, particularly among young users.[75]

[73] *See Meta's threat disruptions*, META TRANSPARENCY CENTER (Aug. 27, 2025), https://transparency.meta.com/metasecurity/threat-reporting (providing reports from their work in detecting and countering security threats on their platform from 2017 to the present). Just their latest report, in the last quarter of 2024, lists attacks from Moldova, India, Iran, Lebanon, and Russia. They also report that GenAI is starting to be used, though perhaps not so effectively at present.

[74] *See* Meynadier, J., Malouff, J.M., Schutte, N.S. *et al.* Relationships Between Social Media Addiction, Social Media Use Metacognitions, Depression, Anxiety, Fear of Missing Out, Loneliness, and Mindfulness. *Int J Ment Health Addiction* (2025).

[75] *See* Jashvini Amirthalingam and Anika Khera, *Understanding Social Media Addiction*, CUREUS 16(10): e72499 (Oct. 27, 2024).

As well, while there are cases of loud *and* harmful online speech, not all loud online speech is harmful. Some of the benign amplified speech is even welcome. For example, Lionel Messi's Instagram post after Argentina won the World Cup in 2022 garnered over 75 million likes.[76] Taylor Swift's announcement of her new album, *The Tortured Poets Department*, received over 15 million likes on Instagram.[77] These are not matters of national security, but simply news that many people enjoy and are interested in. One can imagine any number of similar examples of "loud" but not harmful speech. At the same time, while Taylor Swift has over 283 million followers on Instagram, she also has **ample alternative means for communication**, including, for example, her own webpage (https://www.taylorswift.com/). Generally, it would be hard to argue that individuals who have enormous numbers of "followers" or "friends" on social media would not also have ample alternative means for communication. This is especially true of sports figures, actors and actresses, and musical artists, who can easily command large audiences via multiple speech channels.

What about grassroots movements like #MeToo,[78] important news stories about disasters like hurricanes and tornadoes, or government emergency broadcasts—cases where large speech volume may proliferate from a single source, but may not be harmful? Here, too, it is hard to argue that there would not be ample alternative means for communication. Indeed, for these types of newsworthy events, the ample alternative means function as an important and effective mechanism for vetting quality content, which is valuable in itself. Consider that grassroots movements arise from many individual sources even while some single sources may be more poignant than others.

[76] *See* Image posted by Lionel Messi (@leomessi), INSTAGRAM (Dec. 18, 2022), https://www.instagram.com/p/CmUv48DLvxd/.

[77] *See* Image posted by Taylor Swift (@taylorswift), INSTAGRAM, *All's fair in love and poetry... New album THE TORTURED POETS DEPARTMENT. Out April 19* 🤍(Feb. 4, 2024), https://www.instagram.com/taylorswift/p/C28vsIzO_bL/.

[78] *See*, for example, *#MeToo (hashtag page)*, X, https://x.com/hashtag/MeToo?src=hashtag_click (last visited Oct. 23, 2025).

As the grassroots movement gains momentum, it becomes a newsworthy event that becomes increasingly likely to be covered by traditional or legacy news outlets. As well, these news outlets play a critical role for emergency government broadcasts and climate-related events. Both emergency broadcasts and grassroots movements not only have many potential outlets for dissemination, but this type of information can and should originate from many sources. The video documenting the death of George Floyd,[79] for instance, may not have originated from an official news source, but its dissemination can and, ideally, should still travel through multiple news agencies, including traditional or legacy news channels.

Many news and information outlets functioned and existed long before social media and continue to function and play an important role in information dissemination. Indeed, we benefit from an information ecosystem that encourages and supports communication from a multiplicity of online and offline outlets.[80] Democracy Fund, who studies how to build stronger local information ecosystems that strengthen the nation's civic life, finds that "[h]ealthy news ecosystems are diverse, interconnected, sustainable, and deeply engaged with their communities."[81] While it may seem at times that social media has replaced other means of communication, we argue that social media is but one outlet, and that society is served well when it is not *the only* outlet spreading important news.

[79] *See* POLICEACTIVITY, *Full Bodycam Footage of George Floyd Arrest* (YouTube, Aug. 10, 2020), https://www.youtube.com/watch?v=XkEGGLu_fNU&ab_channel=PoliceActivity.

[80] *See* Avery Forman, *How Newspaper Closures Open the Door to Corporate Crime*, WORKING KNOWLEDGE, HARV. BUS. SCH. (Oct. 8, 2021), https://www.library.hbs.edu/working-knowledge/how-newspaper-closures-open-the-door-to-corporate-crime (discussing research linking the closure of local newspapers with violations of publicly listed companies in the paper's circulation area); Elaine Godfrey, *What We Lost When Gannett Came to Town*, ATLANTIC (Oct. 5, 2021), (https://www.theatlantic.com/politics/archive/2021/10/gannett-local-newspaper-hawk-eye-iowa/619847/) (discussing how the loss of local newspaper in a small town in southeastern Iowa caused the town to feel less connected).

[81] *See What is a News Ecosystem*, DEMOCRACY FUND, https://ecosystems.democracyfund.org/what-is-a-news-ecosystem/ (last visited Oct. 23, 2025).

Given that the infrastructure for ample alternative means exists for newsworthy stories, managing online speech volume would then help to disincentivizing bad actors who are more likely to rely on a single outlet. Indeed, stories that originate from a single source should, at least initially, be viewed with caution. Diversifying and containing the power inherent in speech volume is essential for disincentivizing bad actors, fostering individual autonomy and serving democratic self-governance. Consider that if only one news outlet carries a "news story," one would be rightly suspicious of that story's veracity and value. The same skepticism should apply when a single individual commandeers large volume on a single social media platform. That ability or tactic is ripe for exploitation by bad actors, and we must guard against it. While not all news agencies may meticulously vet their stories, the public can recognize that when speech volume is driven by many independent sources, this pattern is an indicator that the information is more reliable than when speech receives all of its volume, either its origination or its amplification, from fewer sources or from only social media sources. A strong ecosystem of diverse sources is also more difficult to exploit for nefarious gain.

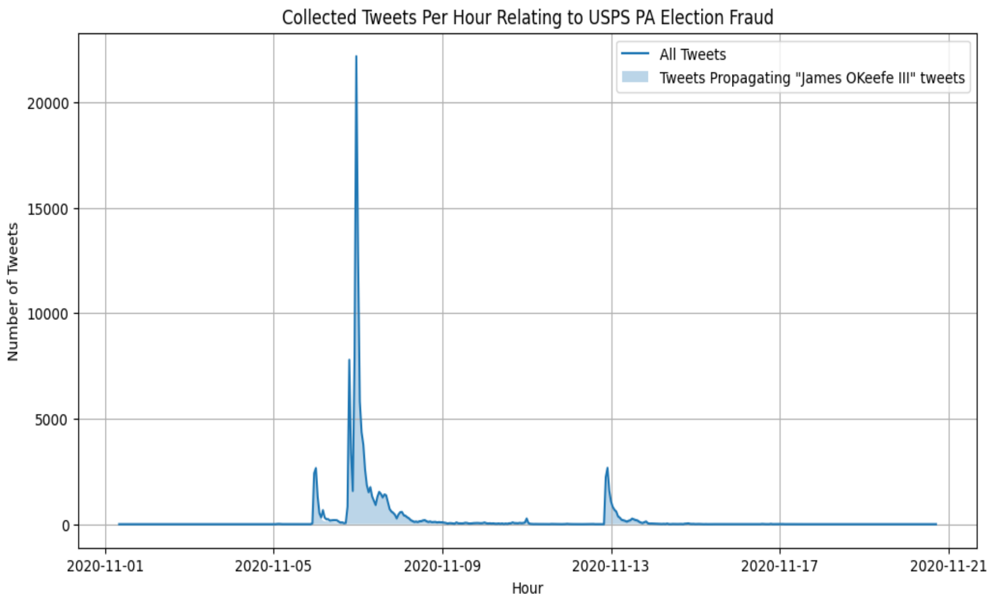

Figure 2. **Graph showing engagement on Twitter with misinformation tweets about the Pennsylvania USPS postmaster back-dating mail-in ballots during the 2020 Presidential Election.**

Consider an event shortly after the 2020 Presidential Election, on November 5, 2020, that offers an example of outsized speech that garnered its attention, not from multiple sources, but from the amplification of a single source on a social media platform. In that incident, one Twitter user, James O'Keefe, posted a video containing allegations that a Pennsylvania USPS postmaster ordered the back-dating of mail-in ballots, so that more could be counted.[82] This story was eventually recanted, including by James O'Keefe, himself, in a tweet on February 5, 2024.[83] Figure 2 (reproduced from Kennedy et al., 2023) shows a temporal graph depicting the tweets per hour relating to the USPS Pennsylvania election fraud narrative.[84] The shaded region reflects the proportion of the total tweets that can be linked to a tweet from @JamesOKeefeIII. As we can see, the engagement with the original content was immediate and substantial. Over 81,000 users interacted with the original tweets from O'Keefe, resulting in more than 122,000 tweets spreading misinformation about election fraud in Pennsylvania. Almost the entire volume of tweets on the subject arises from retweeting a single user. The unfolding of this "news story" is not how one would conceive of a broad scale robust exchange of ideas on a topic of interest.[85] Instead, in the words of Judge Matey, defending this type of algorithmic

[82] James O'Keefe (@JamesOKeefeIII), X (Nov. 5, 2020), https://web.archive.org/web/20201106214000/https://twitter.com/project_veritas (archived capture of now-removed post, which included use of the hashtag #MailFraud).

[83] *See* James O'Keefe (@JamesOKeefeIII), X (Feb. 5, 2024, at 12:46 PM), https://x.com/JamesOKeefeIII/status/1754607379897827598.

[84] *See* Ian Kennedy, Morgan Wack, & Andrew Beers et al., *Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election*. 2 J. QUANT. DESCRIP.: DIGITAL MEDIA (June 2022), at 22.

[85] To be clear, we are proposing an idea and a framework. We are discussing how it applies to a singular piece of content that garners high levels of engagement. Measuring engagement with the number of "likes" or "retweets" is a simple method. Of course, such simplicity invites bad actors to engage in a game of whack-a-mole, perhaps then, instead of liking or retweeting, then creating similar, but not exact content. Implementing our idea requires some technical ingenuity for detecting such patterns. The platforms already engage in this type of "warfare" with tools such as, for example, perceptual hashing algorithms or YouTube's Content ID algorithm. Additional tools can be developed, and we argue likely already exist, for tracing information flow within a network. See, for example, *Ripples*, HINT.FM, http://hint.fm/projects/ripples/ (last visited Oct. 23, 2025) (visualizing information flow on Google); *TrendsMap*, HINT.FM, http://hint.fm/projects/trendsmap/ (last visited Oct. 23, 2025) (showing viral video trends in real-time). As well, KWIC, KWAC, KWOC indexes may be helpful, as may various Natural Language Processing (NLP) tools and Large Language Models (LLMs).

amplification of engaging speech is to "smuggle constitutional conceptions of a 'free trade in ideas' into a digital 'cauldron of illicit loves.'"[86]

To be clear, we are discussing individual pieces of user-generated content, not the broader topics to which those individual pieces of content might contribute. So, if the *topic* were the Super Bowl, and one single tweet generated enormous volume, whether that tweet was true or false, we argue that limiting engagement at some level is warranted for *any single piece* of content regardless of topic/content and without a need to assess its truth value. One can discuss the Super Bowl without the entire discussion centering around a single tweet. If a particular tweet dominates, it is likely either sensationalized or provocative, or it is sufficiently newsworthy to attract attention from ample alternative means. A high-profile example is the deepfake pornography of Taylor Swift that was posted on Twitter. In 17 hours, that particular piece of sensationalized and provocative content garnered 45 million views with hundreds of thousands of likes.[87] There are many such examples of harmful singular pieces of content that receive disproportionate attention from one source. This scenario stands in stark contrast with copious content on a singular subject matter. If many people want to engage with a lot of different content about topics like politics or the Super Bowl, that boisterous, and more likely balanced engagement, produces a far different speech environment than when a single piece of content receives outsized engagement. Free speech principles are exemplified when a large group of individuals gather and are all engaged in conversation together, but when a single individual captures the megaphone and dominates the flow of information, the speech environment deteriorates.

[86] *Anderson v. TikTok*, 116 F.4th 180, 185 (3d Cir. 2024).

[87] *See* Halle Nelson, *Taylor Swift and the Dangers of Deepfake Pornography*, NATIONAL SEXUAL VIOLENCE RESEARCH CENTER (Feb. 7, 2024), https://www.nsvrc.org/blogs/feminism/taylor-swift-and-dangers-deepfake-pornography.

IV. VIABLE TECHNOLOGICAL PATHS FORWARD

To pass constitutional muster, any restriction that manages online speech volume must also be **narrowly tailored**. The narrowly tailored component is assessed in the legal realm but requires the development of a technological solution. Although the algorithms used by the platforms are unknown and evolving, we can gain some insight into how they work from considered the economic interests of the platforms as well as the platforms' own discussion of their algorithms. For instance, in 2021, Mark Zuckerberg publicly shared insights into Meta's research and content moderation strategies.[88] Although his description includes only general parameters and lacks the details that would enable one to reconstruct the underlying algorithms, it nonetheless offers valuable information for our purposes. For instance, for Instagram, Zuckerberg states that "[t]he AI system behind Instagram Search automatically orders search results by predicting what you'll find most *valuable* and *relevant*" (emphasis added).[89] In the description of "how AI delivers content to you," he states that "[t]hese prediction models use underlying input signals to help select content you're most likely to *engage* with" (emphasis added).[90] So, while the exact implementation of these algorithms remains uncertain, we know that engagement and relevance play central roles in determining what content populates a user's social media feed. This is not surprising, as these platforms are commercial enterprises, and "engagement" is the key commodity they aim to capture.[91] Although these statements were made in 2021, and four years of time in the

[88] For Meta's content moderation, see Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (May 5, 2021), https://www.facebook.com/notes/751449002072082/). For Instagram, see *Instagram's Search AI System*, META TRANSPARENCY CENTER (Mar. 7, 2025), https://transparency.meta.com/features/explaining-ranking/ig-search/.

[89] Zuckerberg, *A Blueprint for Content Governance and Enforcement*, *supra* note 82.

[90] *Id*.

[91] Our discussion of engagement assumes a measure of engagement that is not biased toward any particular type of content. How faithfully various platforms adhere to such an unbiased model is unknown. There has been increasing discussion of bias in the Twitter algorithm. *See* Giulio Corsi, *Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study*. 13 EPJ DATA SCI. 18 (2024); Jack Gillum, Alexa Corse, & Adrienne Tong, *X Algorithm Feeds Users Political Content—Whether They Like it or Not*, WALL ST. J. (Oct. 29, 2024), https://www.wsj.com/politics/elections/x-twitter-political-content-election-2024-28f2dadd.

technology space is like an eternity, as rapid innovation and constant disruptions can make even recent advancements feel outdated, the business goals of social media platforms have remained constant. In essence, the more engaged a user is with the content, the more content that user will consume. The more content consumed, the greater the revenue that the platform is able to generate through advertisements, which drives their bottom line. While platforms may have other motivations, it is undeniable that they pursue their business interests by fostering and increasing user engagement.

A. The Dark Side of Engagement

There is a dark side to "engagement," however, which the platforms also realize. As Zuckerberg notes, "[o]ne of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content."[92] Meta is well aware of the potential peril of this proclivity as Zuckerberg discusses how this tendency "can undermine the quality of public discourse and lead to polarization."[93] He provides the graphic shown in Figure 3 to demonstrate his point, explaining that "[o]ur research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that [policy] line, people will engage with it more on average—even when they tell us afterwards they don't like the content."[94] We assume, having no reason to believe otherwise, that this tendency of users to gravitate toward sensationalist and provocative content exists on other platforms as well. We accept, as Zuckerberg implies, that this is just an extension of offline human tendencies and behavior and is "not a new phenomenon. It is widespread on cable news today and has been a staple of tabloids for more than a century."[95] Indeed, tabloid journalism has been studied by a number of scholars, and the public's proclivity toward this type of

[92] Zuckerberg, *A Blueprint for Content Governance and Enforcement*, *supra* note 83.
[93] *Id*.
[94] *Id*.
[95] *Id*.

content is well documented. Tabloids sell because their sensational, exaggerated, and dramatic stories grab our attention. They are often short and easy to read, making them widely accessible to a broader audience. Their entire industry centers around engagement.[96] This offline phenomenon unsurprisingly manifests similarly online.

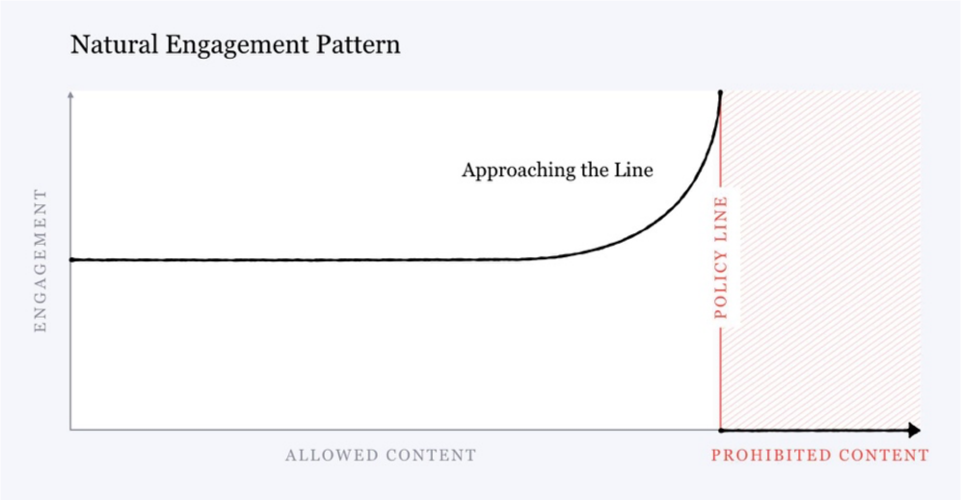

Figure 3. **Meta's description of their research on how its users engage with content more as that content approaches the policy line. Figure reproduced from https://www.facebook.com/notes/751449002072082/.**

Zuckerberg also provides the diagram shown in Figure 4 to explain Meta's response to this human tendency toward sensationalist and provocative content. Basically, as the quality of the content declines, heading toward the crossing of some policy line that separates allowed content from prohibited content, the Meta algorithm penalizes this "borderline content so that it gets less distribution and engagement."[97] The choice to implement their amplification algorithms in this way is made so that users are "disincentivized from creating provocative content that is as close to the line as possible."[98] Although users are unaware of the precise policy line or how close their content may be to it, Meta's assumption is that when users find that their posts are not gaining traction or amassing the number of views that they would like, they learn and

[96] *See, e.g.*, S. Elizabeth Bird, *For Enquiring Minds: A Cultural Study of Supermarket Tabloids*, (Univ. of Tenn. Press, 1992); Henrik Örnebring & Anna Maria Jönsson, *Tabloid Journalism and the Public Sphere: A Historical Perspective on Tabloid Journalism*, 5 JOURNALISM STUD. 283 (2004).
[97] Zuckerberg, *A Blueprint for Content Governance and Enforcement*, *supra* note 83.
[98] *Id.*

adapt to create less sensationalist and provocative content that will receive more distribution and engagement.[99]

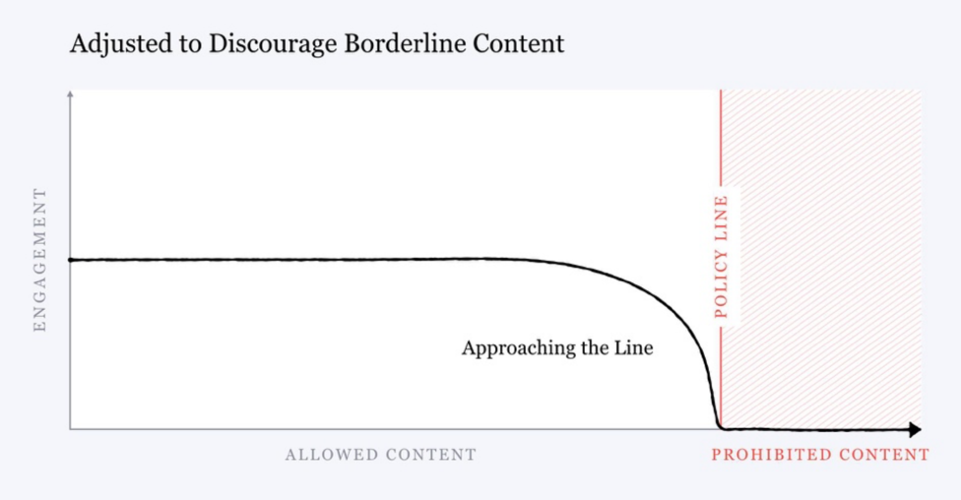

Figure 4. **Meta's description of their policy to penalize content that approaches the policy line by reducing its distribution and engagement. Figure reproduced from https://www.facebook.com/notes/751449002072082/.**

Zuckerberg also claims that their content moderation strategy includes "train[ing] AI systems to detect borderline content so we can distribute that content less" and that once they are able to achieve the goal of creating an AI system that detects borderline content well, they can "proactively remove harmful content and reduce the distribution of borderline content."[100] Note that Zuckerberg's emphasis is on determining *content* and thus necessarily content-based.[101] At first blush, this policy and these strategies toward reducing harmful content all seem to make sense since all of the discussion about the problems with social media platforms have centered around reducing harmful content. Accordingly, Meta is training AI tools to determine where on the horizontal *x*-axis in Figure 4 that measures "good"/allowed and "bad"/prohibited content a particular post

[99] *See id.*

[100] *Id.*

[101] *See id.* ("[O]ur work fighting misinformation includes . . . proactively identifying fake accounts, which are the source of much of the spam, misinformation, and coordinated information campaigns. This approach works across all our services, including encrypted services like WhatsApp, because it focuses on patterns of activity rather than the content itself."); *see also* Mark Zuckerberg, *Preparing for Elections*, FACEBOOK (Mar. 13, 2021), https://www.facebook.com/notes/737729700291613/ (discussing behavior-based, rather than content-based, moderation strategies).

should be placed, and whether that post is on the allowed side or the prohibited side of their policies. They are assessing content in order to determine how close a particular piece of content is to their "policy line."[102]

B. A Content-Neutral Proposal

We note, however, that an easier and *content-neutral* method for achieving the same outcome is to focus on the measure of engagement shown on the vertical *y*-axis. Consider the marked-up graphic shown in Figure 5. In order for Meta to place content along the horizontal *x*-axis, it needs to train an AI system to read and understand content (a difficult task) and then to place that content somewhere on the continuum of "allowed content" to "prohibited content" (a task that requires human-infused values). To determine if content falls in the prohibited content bucket, some hard cutoff for a "policy line" must be drawn. There is no known scientifically principled way to draw this policy line, and the exact process by which Meta identifies it remains unknown. Training AI systems to detect content is quite a difficult task and neither Meta nor any other entity has successfully accomplished it. Arguably, the task is not objectively accomplishable since drawing this policy line is necessarily a value-laden judgment, and not something that technology could achieve without incorporating human-infused values.

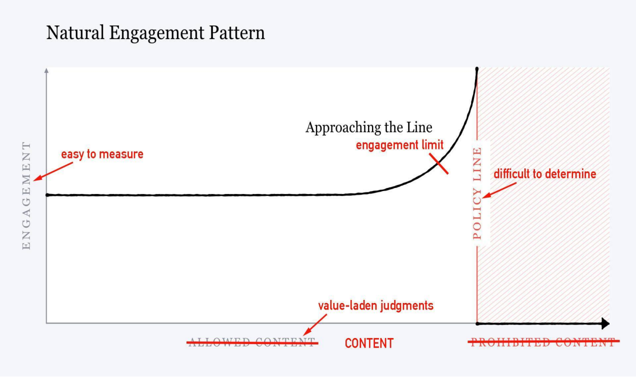

Figure 5. **Our proposal on how to re-interpret Meta's research on content engagement and moderation. Figure modified from https://www.facebook.com/notes/751449002072082/.**

[102] Mark Zuckerberg, *Illustration of Meta's Borderline Content Distribution Curve* in *A Blueprint for Content Governance and Enforcement*, Facebook (May 5, 2021),https://www.facebook.com/notes/751449002072082

Instead of making value-laden judgments about whether a particular piece of content is allowed or not allowed, and instead of drawing some policy line that is difficult to determine or define, even if we could agree on the human values that should inform that choice, platforms can alternatively just set an engagement limit for all content (with no knowledge whatsoever of what that content might be) since we know *ex ante* that, according to Zuckerberg, himself, as "piece of content gets close to that [policy] line, people will engage with it more on average." [103] The technological tools to measure engagement already exist and are part and parcel of their entire business strategy to engage users with the content of their platforms. Meta already employs mechanisms to limit engagement for particular content since they told us that they do so for content that they deem is close to the policy line. Their vested interest here is clear and evident. According to a Facebook engineer, Meta is continually monitoring and tweaking their engagement algorithms.[104] Twitter as well engages in this type of content curation: "[r]estricting the reach of Tweets, also known as visibility filtering, is one of our existing enforcement actions that allows us to move beyond the binary 'leave up versus take down' approach to content moderation."[105] Moreover, and critically, engagement can be measured objectively. There is no need to entangle ourselves with the difficult and value-laden task of identifying "good" and "bad" content because we already have an exceptionally good proxy measure in engagement.

Meta's articulated approach is content-based and much more complex to execute that our proposed content-neutral approach since their approach requires significant

[103] Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, Facebook (May 5, 2021), https://www.facebook.com/notes/751449002072082.

[104] *See* Krishna Gade (@krishnagade), X (Feb. 11, 2021, at 8:55 AM PT), https://x.com/krishnagade/status/1359908897998315521.

[105] *See Freedom of Speech, Not Reach: An Update on Our Enforcement Philosophy*, X: BLOG (Apr. 17, 2023), https://blog.x.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy.

technological advances that require human-infused values.[106] Determining how to moderate content without making value-laden judgments is not only perplexing, but impossible. To be clear, we are not claiming that their research is not useful or that they may not be within their rights to moderate their platform as they wish. Surely, given the mind-numbing amount of content being posted on social media platforms, it is helpful for platforms to develop new "AI systems" for identifying harmful content. These systems can help identify harmful content quickly and proactively so that they can be moderated before engagement measures would even issue warnings. In 2018 during Congressional testimony, Mark Zuckerberg claimed that "99% of the ISIS and al-Qaeda that we take down off of Facebook, our AI systems flag before any human sees it." Whether that number is accurate or not, that new AI systems would be, or are, beneficial does not preclude a multi-pronged approach or negate the value of our content-neutral proposal. We must also realize that progress in this sphere also entails switching from this mindset of needing to screen all content in some value-laden way in order to assess its harm to, instead, considering and highlighting the harms emanating from how platforms moderate, guide, and control online speech volume.

To be sure, Meta's approach to content moderation is evolving. Their most recent announcement seems to be a significant departure from their previous policies. The latest announced policy has three key takeaways. Quoting from their press release:[107]

> - Starting in the US, we are ending our third party fact-checking program and moving to a Community Notes model
> - We will allow more speech by lifting restrictions on some topics that are part of mainstream discourse and focusing our enforcement on illegal and high-severity violations
> - We will take a more personalized approach to political content, so that people who want to see more of it in their feeds can

[106] This mechanism is employed by X as well, see *id.* X demotes content that it claims violates their rules. X attaches "publicly visible labels to Tweets identified as potentially violating our policies letting you know we've limited their visibility", but how X determines that a rule has been violated is unclear from its descriptions. *Id.*

[107] Kaplan, *More Speech and Fewer Mistakes*, *supra* note 21.

The impact of this announcement will unfold over time. However, Zuckerberg acknowledges that the reality of the new policies will be a "trade-off." [108]Since they will moderate less and take fewer things down, that "means that we're going to catch less bad stuff."[109] As we have discussed, and they have admitted, the "bad stuff" generally garners high levels of engagement, so they are, in effect, announcing that their "policy line" (from Figure 3) is moving further to the right. [110]That is, a large swath of what was previously "prohibited content" will now be "allowed content." Moreover, going forward, they will "focus these systems on tackling illegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams."[111] In other words, they will continue to penalize content, but will penalize less content. In light of these policy changes, all of our arguments remain, and, indeed, become even more vital and pressing. To be sure, while the platforms will moderate less, they will continue to measure engagement to predict what users will find most valuable and relevant, and that content will include more engaging "bad stuff."

V. DISCUSSION

Although not introduced with much fanfare, the advertisement funding model for social media was utterly revolutionary. Google launched Google AdWords in 2000. Mark Zuckerberg unveiled Facebook Ads in 2007.[112] In Fiscal Year 2023, the major social media platforms reported *billions* in revenue. Meta posted $134.9 billion in revenue, with

[108] Mark Zuckerberg, *Illustration of Meta's Borderline Content Distribution Curve* in *A Blueprint for Content Governance and Enforcement*, Facebook (May 5, 2021), https://www.facebook.com/notes/751449002072082.

[109] Video posted by Mark Zuckerberg, META, *It's time to get back to our roots around free expression. We're replacing fact checkers with Community Notes, simplifying our policies and focusing on reducing mistakes. Looking forward to this next chapter*, (Jan. 7, 2025), https://www.facebook.com/watch/?v=1525382954801931.

[110] "Mark Zuckerberg, *Video Announcing Model Changes*, Facebook, https://www.facebook.com/watch/?v=1525382954801931 (last visited Oct 19, 2025) (stating "bad stuff" at 2:59)

[111] Kaplan, *More Speech and Fewer Mistakes*, *supra* note 21.

[112] *See Facebook Unveils Facebook Ads*, META (Nov. 6, 2007), https://about.fb.com/news/2007/11/facebook-unveils-facebook-ads/.

99% of their total revenue coming in from ads.[113] Alphabet, the parent company of Google and YouTube, reported revenue of $307 billion with about three quarters of its revenue coming from advertisements.[114] Reddit reported $804 million with 98% of that coming from ad revenue.[115] Twitter revenue came in at $3.4 billion where over 70% was from advertising.[116] TikTok, which saw $120 billion in revenue, has a more diverse revenue source with, for instance, an in-app marketplace that also generates income, but still brought in billions from ads.[117] The social media industry lives and dies on advertising revenue, which is strongly tied to the engagement of their users on their platform.[118]

The social media platforms have amassed billions in paid advertisements because their users spend an average of 2.5 hours on social media alone *every day*. Compare this with the average user spending 30 minutes a *month* on the Internet in 1996 (and not on social media, which did not exist back then).[119] Consider as well that the social media platforms are in competition with one another. They are all vying for the same thing—more of your time spent on *their* platform because that is how they increase the advertisement revenue that funds their businesses. It is a zero-sum game, where the commodity is your attention. The key technological development for competing in this

[113] Meta Platforms, Inc., Annual Report (Form 10-K) for the Fiscal Year Ended Dec. 31, 2023, at 91, https://www.sec.gov/Archives/edgar/data/1326801/000132680124000012/meta-20231231.htm#ibbcdb9a98fd34a92b3add929872a8009_88.

[114] Press release from Alphabet Inc., Alphabet Announces Fourth Quarter and Fiscal Year 2023 Results (Jan. 30, 2024), https://s206.q4cdn.com/479360582/files/doc_financials/2023/q4/2023q4-alphabet-earnings-release.pdf.

[115] Reddit, Inc., Registration Statement (Form S-1) (filed Feb. 22, 2024), https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm.

[116] Twitter, Inc., Quarterly Report (Form 10-Q) for the Period Ended Mar. 31, 2022 (filed May 2, 2022), https://www.sec.gov/Archives/edgar/data/1418091/000141809122000075/twtr-20220331.htm.

[117] Reuters, *TikTok's US revenue hits $16 bln as Washington threatens ban, FT reports*, Mar. 15, 2024, https://www.reuters.com/technology/tiktoks-us-revenue-hits-16-bln-washington-threatens-ban-ft-reports-2024-03-15/ (US revenue was $16 billion. The $120 billion refers to China's ByteDance revenue.)

[118] Our discussion does not include other platforms, like Medium or the *Wall Street Journal*, which have subscription-based models, though may also have advertising revenue. We do not purport to draw a strong line here. A policy about where this line might be may become an issue, but the large social media platforms are quite clearly over any reasonable standard.

[119] *Compare* Ricky Ribeiro, *The Internet: 1996 vs. 2011*, BIZTECH (May 31, 2012), https://biztechmagazine.com/article/2012/05/internet-1996-vs-2011-infographic, with Simon Kemp, *The Time We Spend on Social Media*, DATAREPORTAL (Jan. 31, 2024), https://datareportal.com/reports/digital-2024-deep-dive-the-time-we-spend-on-social-media.

information ecosystem is the recommendation algorithm. Every social media platform has such an algorithm: Facebook ads, Twitter ads, Twitter Trending Topics, TikTok For You feeds, Instagram Suggested Posts, recommendation videos on YouTube, and Reddit Popular Communities. The platforms are in competition with one another for your attention—the longer you stay on Facebook, the less time you will have for TikTok. Commanding your attention to raise their advertising revenue lies at the heart of how social media platforms determine which content proliferates on their platforms.

Understanding and accepting that uneven speech volume online is problematic and can be addressed within First Amendment constraints is an important first step. However, redirecting the trajectory of social media discourse is not simple. How one might structure regulation of online speech volume to mitigate its harms remains a thorny and unresolved problem. Although redesigning recommendation systems to temper the distortions of engagement-maximization is a complex challenge, the platforms are uniquely positioned to evaluate how algorithmic adjustments could elevate societal well-being while still incorporating engagement as one factor rather than the overriding objective of their algorithms. Nevertheless, identifying a viable legal pathway for regulation matters precisely because a credible prospect of regulatory intervention can realign platform incentives that, in the absence of such pressure, remain firmly tethered to the status quo of engagement-driven profits.

At present, the platforms have incentives both to develop new "AI systems" as well as to moderate harmful online content. These dual incentives create a delicate balancing act for them as they aim to keep their users engaged while avoiding associations with highly controversial or explicit content. The platforms want to maintain user attention without allowing their content to devolve into excessive toxicity. The new "AI systems" at Meta are likely to focus on particular types of content that violate their Community Standards. This includes, for example, violence and incitement, suicide and self-injury, human exploitation, and sexual solicitation.[120] However, the platforms are not likely to simply volunteer to make changes that reduce engagement on their platforms since that runs so directly afoul of their profit motives and business model. Instead, they are likely to continue to focus their efforts on content-based "AI systems" that reduce particular forms of harmful content.

Importantly, while developing these "AI systems" and reducing the role of engagement in the recommendation algorithms would both be helpful in moderating harmful content, the two are wholly different strategies with markedly distinct consequences. Because of the distinct consequences, social media platforms are likely, on their own, to develop new "AI systems" but are unlikely, on their own, to reduce the focus on engagement in their recommendation algorithms. As we have already noted, Meta purported to reduce engagement for content that was close to their "policy line." So, at some level, they already moderate engaging content, which we do not dispute, but their actions here are not based on engagement. The difference is subtle, but important. Meta *first* assesses content. *If* that content is deemed harmful, it reduces its amplification. If Meta is able to, *ex ante*, identify harmful content well, then there would be little left to discuss about content moderation on social media platforms. However, we know that this strategy, even if well-meaning and actively practiced, is insufficient. This strategy also, auspiciously for the platforms, leaves the engagement portion of their algorithms intact.

We are suggesting a different strategy that monitors engagement (not content) and manages that engagement (regardless of content) because we recognize that *uneven and large amplification volume, itself, is the harm*. Meta's strategy of assessing content is market-driven. Since their users do not want to be on a platform overrun by, say, hate speech or pornography, Meta moderates this type of harmful content. However, changing their

[120] *See Policies*, META: TRANSPARENCY CENTER, https://transparency.meta.com/policies/ (last visited Oct. 19, 2025) (providing the full list of policies and explanations).

recommendations algorithms so that they are less focused on increasing engagement generally is a completely different strategy that is not market driven. While diminishing the role of engagement metrics in their algorithms would have societal benefits, it also, by definition, reduces user engagement with their platform. If one platform chose to prioritize societal interests by intentionally downplaying sensational content, users would find that platform less engaging and flock to a competitor that embraces it, thereby undercutting the ethically responsible platform's market share. This conflict illustrates why the platforms are not likely to change course on their own and why regulatory intervention is necessary.[121] Based on their own internal research, the platforms may be able to identify creative strategies that deliver engaging quality content, but they need external incentives to redirect their research in that direction.[122] At present, the economic gain from amplification strategies is known, substantial, and phenomenally successful. But we know that that success comes at the cost of increasing societal harm.

Meta's newly announced content moderation strategies, and arguably how Twitter has been managed for some time, yet further bolster their profitability and further increase societal harm. The platforms are plainly grandstanding when they claim that their motivation is to champion free speech since their policies do not necessarily advance that ideal in a meaningful way—simply more content or more people "speaking" does not foster open discourse when this larger corpus of voices is fed though amplification algorithms that funnel based on engagement. When these algorithms disproportionately amplify sensationalist, divisive, emotionally charged, "engaging content," they distort

[121] *See* Jeff Horwitz, *Facebook Executives Shut Down Efforts to Make the Site Less Divisive*, WALL ST. J. (May 26, 2020), https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499; *see also* Karen Hao, *The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why*. MIT TECH. REV. (Oct. 5, 2021), https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/; Ryan Mac & Cecilia Kang, *Whistle-Blower Says Facebook 'Chooses Profits over Safety,'* NY TIMES (Oct. 3, 2021), https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html.

[122] *See* Jeff Allen, *Why Is Instagram Search More Harmful Than Google Search?*, INTEGRITY INSTITUTE (Feb. 13, 2024), https://integrityinstitute.org/blog/why-is-instagram-search-more-harmful-than-google-search.

the underlying speech in harmful ways rather than facilitating genuine dialogue. Without modifying the role of engagement in their amplification algorithms, their announced content moderation strategies, rather than moving society toward democratic self-governance, instead pushes public discourse further into manipulated fragmentation.

Social media has changed the *manner* in which speech now manifests. Its algorithmic curation has expanded the reach of speech, but also transformed its content, structure, and dynamics. Although social media has enabled more speech, their algorithms also control the speech environment. One pathway for regulation could thus focus on restoring individual autonomy, a foundational principle of democratic self-governance and a core rationale for robust First Amendment protections. While algorithmic recommendation systems ostensibly help individuals discover relevant content, they also make substantive choices on users' behalf, narrowing exposure and diminishing user autonomy in determining what speech to engage with and what communities to associate with. Regulation grounded in user autonomy could require platforms to adopt "algorithmic choice" mechanisms. For example, every major platform at the moment, with the except of TikTok, allows its users to choose a chronological feed rather than the default recommendation algorithm. However, these chronological feeds are often not obviously available and switching the algorithm that governs the content in a user's social media feed is a non-trivial task for many users. It would be simple, however, for the platforms to make the chronological feed the default and then require users to affirmatively opt-in to their recommendation systems, if they so wished. This would alter the manner in which speech appears as well as decrease the platform's control over the speech encountered by its users.

Another option that would neutralize the adverse impact of engagement-based algorithms might be to require that users generally have greater control over their news feeds. Bluesky, a social media platform that opened to the public in February 2024, and spiked its user base to over 25 million following the 2024 Presidential election, has experimented with some unique features along these lines, including My Feeds, a model that allows its users to choose from a variety of algorithms to power their social feed.[123] Bluesky uses the AT protocol, which is an open Application Programming Interface (API) that allows outside developers to communicate, interact with, and gather data from Bluesky. This functionality makes it possible for middleware to be developed.[124] Middleware is third-party software that functions as an intermediary layer between users and platforms. At present, the social media platforms are the intermediaries between a user and the platform, but middleware would decentralize and diversify the intermediary role. This type of decentralization would also functionally increase individual autonomy in the social media ecosystem.

The AT protocol that underlies the Bluesky platform permits flexibility beyond algorithmic choice. Potentially, it enables the creation of a decentralized platform that can be hosted on any server, further allowing an even larger degree of user customization and even greater user control over their information environment.[125] Mastodon has similar features that allow this more open development environment by operating with the ActivityPub protocol. More ambitious regulation could explore incentives for the decentralization of social media infrastructures, such as encouraging adoption of the ActivityPub protocol that underpins the Fediverse.[126] The Fediverse is a collection of independent social media platforms that are able to interact with one another, where each independent platform is able to design their own rules.[127] So if Facebook were part of

[123] *See* Chris Stokel-Walker, *Bluesky's Custom Algorithms Could Be the Future of Social Media*, WIRED (June 3, 2023), https://www.wired.com/story/bluesky-my-feeds-custom-algorithms/ (describing Bluesky's "My Feeds" feature that lets users choose the algorithm powering their feed).

[124] Francis Fukuyama, Richard Reisman, Daphne Keller, Aviv Ovadya, Luke Thorburn, Jonathan Stray, and Shubhi Mathur, *Shaping the Future of Social Media with Middleware* (Dec. 2024), FOUNDATION FOR AMERICAN INNOVATION.

[125] *See The AT Protocol*, BLUESKY: DOCS, https://docs.bsky.app/docs/advanced-guides/atproto (last visited Oct. 19, 2025) (introducing the protocol and its federated, portable-identity design).

[126] https://en.wikipedia.org/wiki/ActivityPub

[127] https://en.wikipedia.org/wiki/Fediverse

the Fediverse, one could interact with Facebook without being on the Facebook platform itself. Although this may sound like a foreign and difficult concept to implement and interact with, note that Bluesky has already demonstrated that a federated social media platform can look and feel like Twitter. Moreover, Meta's Threads, which is built on the ActivityPub protocol, has also demonstrated the technological viability of federated social networking models that retain the friendliness and usability of traditional platforms.

These types of regulations are aimed at changing the manner in which speech appears on social media and restoring individual autonomy to its users. At the same time, enacting such regulations is non-trivial since Congress moves slowly and the courts are even slower at enacting change while technology blazes forward. Accordingly, the fastest and most effective reforms are likely those that are aimed directly at incentivizing the platforms to move away from their engagement maximizing algorithms and to innovate in new directions. On this path, Lessig has proposed an engagement tax on advertising profits that would increase exponentially as platform users exceed some baseline "healthy" level of engagement.[128] Since the current core drive of their revenue model is increasing engagement through their recommendation algorithms, reducing the amount of revenue the platforms could derive in this way would certainly change their business model and thus the design of their algorithms that control the manifestation of online speech. Similarly, Acemoglu and Johnson advocate a digital advertising tax that imposes a flat tax of 50% on annual digital advertising revenue exceeding $500 million.[129] Though these two proposals differ in their tax structures, they share the same premise:

[128] *See* Ivey-Elise Ivey, *What is AI Doing to America's Democracy? – The London School of Economics Phelan Centre Event Review*, LOND. SCH. ECON. (Oct. 26, 2024), https://blogs.lse.ac.uk/usappblog/2024/10/26/what-is-ai-doing-to-americas-democracy-lse-phelan-centre-event-review/ (summarizing remarks by Lawrence Lessig proposing an "engagement tax" on social-media advertising profits that increases as user engagement exceeds baseline).

[129] Daron Acemoglu and Simon Johnson, *The Urgent Need to Tax Digital Advertising*, NETWORK L. REV. (Spr. 2024).

meaningful regulation needs to reshape the underlying business model by making user engagement less profitable. Once financial incentives shift, the platforms are also incentivized to innovate in new directions that are hopefully are more beneficial for society and less exploitative of individual weaknesses.

It is possible, and even desirable, to alter financial incentives and impose measures that strengthen individual autonomy. Doing so regulates the distribution and architecture of speech rather than its content, thereby addressing the systemic harms of viral amplification without infringing on individual expression or the platform's editorial choices. As such, they offer promising avenues for mitigating the societal harms associated with contemporary social media systems. Whether social media algorithms constitute protected platform speech is the subject of ongoing legal battles, but even if algorithmic editorial curation is the platform's protected speech, a well constructed regulation or time, place, and manner restriction remains a legally viable path.

CONCLUSION

Lawrence Lessing writes that a

> code of cyberspace, defining the freedoms and controls of cyberspace, will be built. About that there can be no debate. But by whom, and with what values? That is the only choice we have left to make.[130]

At present, the values that are defining the freedoms and controls of cyberspace are being driven and shaped by the large social media platforms. Ideally, we continue to leave all such choices to the platforms. They would continue to innovate, guided by the strong incentives to moderate content on their platforms. As we have discussed, they are motivated by a general commitment to American free speech norms, a commitment to corporate responsibility to positively impact society and adhere to ethical considerations,

[130] LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE 6 (2006).

and their own business interests. And they clearly act on those incentives. The grand hope is that these incentives will compel the platforms to self-govern well, and so, decisions about how the platforms operate will remain in the domain of these autonomous private actors. They would be free to innovate and develop without state interference or constraints.

Social media platforms have immense power to control, shape, and even manipulate public discourse. They have the potential to expand how free speech is realized for all by providing a unique medium for the robust exchange of ideas.[131] Realizing this ideal world requires a host of complex decisions to align correctly. One critical decision involves how platforms weigh their own economic interests against societal interests. When these two interests are not in conflict, the decision is easy. When they are in conflict, there needs to be some type of framework in place that incorporates or even prioritizes societal interests over business profit. That framework is a regulatory framework. Any regulatory approach must consider compliance with existing laws. We have proposed some viable paths forward that would institute safeguards to keep the balance from tipping too far in favor of financial pursuits.

As with any industry, government interference is never preferred, but may be necessary. Congress, and perhaps the judiciary, need to step in when destabilizing forces emerge that threaten a coherent vision of democratic discourse, which is likely to occur when incentives become misaligned. Proper regulation in the social media space allows us to retain and nurture the innovation of these platforms that has allowed all of mankind to now be seamlessly connected to one another. This connection can inspire the human collective to its best self, or it can magnify our worst tendencies.

[131] *But see* Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224, 260 (2011) (arguing that aspects of internet design and practice can make users less free).

Regulation targeting the recommendation algorithms that control online speech volume would help curb harmful amplification and strengthen the individual autonomy that ultimately serves democratic self-governance, the foundational purpose of our free speech protections. Our proposal is content-neutral and does not require value-laden judgments by either the government or the platforms. It also does not reduce speech. User-generated content remains, but how and whether it is amplified or not hews closer to societal interests and not just economic interests. In this way, we protect and harness the benefits of online speech while mitigating the attendant harms.

As Barlow observes, "information is obviously not a thing. In fact, it is something that happens in the field of interaction between minds or objects or other pieces of information."[132] Today, we find ourselves in an attention economy where online intermediaries control not only the visibility of individual pieces of information but also how they interact. The design of these platforms shapes our new information ecosystem, governing the flow of information and influencing how conversations are framed and unfold. The battlefield of the mind has shifted fundamentally with the rise of modern technology, reshaping how we perceive, process, and respond to information. As we navigate this new information age, we must ensure that the code of cyberspace and the continued development and innovation on the Internet is grounded in the human values we cherish and not unduly driven by the profit motives of the social media industry.

[132] John Perry Barlow, *The Economy of Ideas: A Framework for Patents and Copyrights in the Digital Age (Everything You Know About Intellectual Property Is Wrong)*, WIRED (Mar. 1, 1994), https://www.wired.com/1994/03/economy-ideas/.