SOCIAL MEDIA, DIGITAL CENSORSHIP & THE FUTURE OF DEMOCRACY

By

Kalev Hannes Leetaru

SOCIAL MEDIA, DIGITAL CENSORSHIP & THE FUTURE OF DEMOCRACY

Two decades ago, President Bill Clinton predicted that the Internet would help democratize the world, famously quipping that China's early efforts to censor the web were "like trying to nail Jell-O to the wall." [1] Just over a decade later, Google's Chief Legal Officer David Drummond noted that "governments have learned in what might be the steepest learning curve in history that they can shape this global phenomenon called the Internet and in ways that often go beyond what they can do in the physical world and they're doing so at an alarming pace." [2] In just a few years, the web's utopian dream of free speech and democracy for all had given way to an Orwellian world of censorship and surveillance. Today the future of social media censorship represents nothing less than the existential battle over the future of democracy itself. [3]

How did we arrive at a point in history in which a handful of unelected billionaires wield near-absolute control over digital speech, with the power to censor citizens and governments alike, arbitrate "acceptable speech" for the entire planet, determine "truth" and even silence the president? How did the companies that once refused to silence terrorists [4] and held "free speech" as an absolute right [5] devolve into global censorship machines whose reach increasingly extends into the physical world?

America's two centuries of experiments attempting to balance freedom of expression with the desire to constrain "harmful" speech demonstrates the sheer impossibility of devising a consensus view of "acceptable speech" that works for a diverse nation. The unintended consequences of the myriad approaches the nation has explored, from local to federal, government to private, mandatory to voluntary, courts to capitalism, reminds us that anything short of unfettered speech becomes an "intractable" problem that inevitably silences the very underrepresented voices they were designed to empower. Is there any hope?

How are modern social media platforms similar to past technologies from the post office to motion pictures to broadcasters to cable monopolies and how were those affordances addressed by the legal and societal frameworks of the era? With what new challenges do social platforms confront society?

Finally, what are concrete steps, from legal to educational to technical that policymakers can take to confront these challenges?

[1] https://www.nytimes.com/2000/03/09/world/clinton-s-words-on-china-trade-is-the-smart-thing.html
[2] https://www.youtube.com/watch?v=hJz4V3E5ea4
[3] https://www.dni.gov/files/ODNI/documents/assessments/GlobalTrends_2040.pdf
[4] http://www.nytimes.com/2015/03/25/world/middleeast/behind-a-veil-of-anonymity-online-vigilantes-battle-the-islamic-state.html
[5] https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech

Social media … makes it easier for people to affiliate with others around the world who share common characteristics, views, and beliefs. [Platforms] create echo chambers of like-minded users who share information that confirms their existing worldviews and limits their understanding of alternative perspectives. Over time, this dynamic is increasing awareness of and building new connections between previously isolated groups, while also polarizing people's perceptions of policies, public institutions, events, moral issues, and societal trends. Such polarization will lead to a proliferation of competing, entrenched perspectives, limiting opportunities for compromise and decreasing societal cohesion. During the next 20 years, the algorithms and social media platforms that curate and distill massive amounts of data will produce content that could overtake expertise in shaping the political and social effects engendered by a hyperconnected information environment. Power increasingly will be wielded by the generators of content as well as the arbiters of who gets to see it. Social media platforms will reinforce identity groups, or foster new and unanticipated groupings, and accelerate and amplify natural tendencies to associate with people who share the same views, often engendering competing visions of the truth about an issue. The platforms will make it easier for competing opinion leaders - including from marginalized groups - to publish their views and debate among themselves, honing the cohesiveness and "market appeal" of their messages. This effect is magnified because people rely on their own identity communities for information and piggyback on the knowledge of others. People will also use social identities such as culture, ethnicity, nationality, and religion as critical filters for managing information overload, potentially further fragmenting national identities and undermining trust in government. These identities provide a sense of belonging and reinforce norms about how group members should behave, rules about whom to trust, and beliefs about complex issues. Identity-based violence, including hate and political crimes, may increasingly be facilitated by social media. In India, social media and mobile messaging platforms have become a key force behind viral falsehoods, such as rumors that quickly spread among some Hindus regarding Muslims' alleged slaughter of cows or possession of beef, which led to the "cow vigilante" lynching of Muslims. Publics increasingly will depend on their favorite gatekeepers - such as news media outlets, social media platforms, and trusted voices of authority - to sift truth from fiction. Efforts to arbitrate controversial content, such as flagging or removing demonstrably false claims, are unlikely to be effective in changing beliefs and values aligned with one's closely held identities, however. Identity-based beliefs tend to eclipse truth-seeking because of the overriding need to belong, obtain status, understand the social world, maintain dignity, and feel morally justified. … [By 2040] polarized societies, shaped by social media, [will lead] to more political deadlock and wild policy swings. These factionalized communities, primarily in democratic countries, [will be] unable to take effective action on the economy, the environment, migration, and foreign policy.

Global Trends 2040, United States National Intelligence Council [6]

[6] https://www.dni.gov/files/ODNI/documents/assessments/GlobalTrends_2040.pdf

THE PROMISE OF AN UNFETTERED WEB

The World Wide Web was seen in its earliest days as nothing short of the greatest democratizing force ever created. It was heralded as "the first truly 'mass' medium" in that "while few individuals and groups can publish books or newspapers, make a film, or produce a radio or television program, any person with a personal computer and a modem can communicate" with the entire planet. [7] Its decentralized and ephemeral nature meant the strict speech rules and government oversight of previous centralized communications technologies did not readily apply, while its global nature meant a single voice could now break free of local publishing rules and be heard by the world, regardless of its views.

This sudden global empowerment meant that the disenfranchised, underrepresented and silenced voices of the world could now speak out to question established societal order. This "potential for increasing the political participation of the disenfranchised" made the nascent web an urgent censorship target for governments across the world. [8]

Even in these early days, the future of the web was described as an existential battle between silencing violent, illegal and hateful speech and creating an unfettered unmoderated zone of free expression. Human Rights Watch was among those of the era to demand in 1996 that governments "repudiate the international trend toward censorship and to express unequivocal support for free expression guarantees on-line," noting that: [9]

> *Governments around the world, claiming they want to protect children, thwart terrorists and silence racists and hate mongers, are rushing to eradicate freedom of expression on the Internet … Restrictions on Internet access and content are increasing worldwide, under all forms of government. Censorship legislation was recently enacted in the United States, the birthplace of the Bill of Rights as well as of this new communications medium… Authoritarian regimes are attempting to reconcile their eagerness to reap the economic benefits of Internet access with maintaining control over the flow of information inside their borders. Censorship efforts in the U.S. and Germany lend support to those in China, Singapore, and Iran, where censors target not only sexually explicit material and hate speech but also pro-democracy discussions and human rights education. Proposals to censor the Internet wherever they originate violate the free speech guarantees enshrined in democratic constitutions and international law. In the attempt to enforce them, open societies will become increasingly repressive and closed societies will find new opportunity to chill political expression.*

This dream of bringing the American vision of near-absolute freedom of speech to the world has given way to a far more Orwellian reality. As a handful of unelected billionaires declare sovereignty over the digital world [10] and social media companies increasingly exert control over the societal debates of the world, nothing short of the future of democracy is at stake. How did social media, once seen as such a democratizing force that Twitter postponed a maintenance outage at the White House's request to permit

[7] https://archive.nytimes.com/www.nytimes.com/library/cyber/week/0910hrw.html
[8] https://archive.nytimes.com/www.nytimes.com/library/cyber/week/0910hrw.html
[9] https://archive.nytimes.com/www.nytimes.com/library/cyber/week/0910hrw.html
[10] https://www.facebook.com/zuck/posts/10112681480907401

Iranians to protest a disputed election, [11] turn into the Orwellian world described by the US National Intelligence Council in the quote that opens this report?

Twitter once touted itself as "the free speech wing of the free speech party" [12] and rebuked Congress' calls for it to ban terrorists, proclaiming that "the ability of users to share freely their views - including views that many people may disagree with or find abhorrent" [13] was at the center of its corporate mission. Indeed, most of the early social platforms emphasized unfettered speech above all other considerations. [14] Over the years, this utopian dream has given way to an emphasis on "healthy conversation" [15] and ever-changing [16] enforcement. [17] Facebook today openly muses about what it sees as its corporate responsibility to defend the "norms underpinning democracy" by determining what counts as "free expression" and openly asks questions like "what do we do when a movement is authentic, coordinated through grassroots or authentic means, but is inherently harmful?" [18] Private companies now view their responsibility as being nothing less than shaping the course of the national debate and deciding for themselves what views are "harmful" for society.

Yet for most of their existence, social media platforms have largely avoided censoring elected officials in the U.S. even as they have deleted [19] the accounts of foreign leaders. [20] That all changed as Silicon Valley began labeling President Trump's tweets as "disputed" and "false." [21] As progressive segments of the public embraced this new censorship, platforms moved from merely fact-checking posts to deleting them entirely [22] and threatening to ban other lawmakers. [23]

Before his ban, the courts repeatedly ruled that Trump's Twitter account was an official government outlet and thus he was prohibited from blocking users with whom he disagreed. [24] How then is a private company able to establish "acceptable speech" rules for a government publication or silence it entirely? This represents an unprecedented weakening of political speech: in the broadcast era, radio and television

[11] https://www.reuters.com/article/us-iran-election-twitter-usa/u-s-state-department-speaks-to-twitter-over-iran-idUSWBT01137420090616

[12] https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech

[13] http://www.nytimes.com/2015/03/25/world/middleeast/behind-a-veil-of-anonymity-online-vigilantes-battle-the-islamic-state.html

[14] https://www.realclearpolitics.com/articles/2020/06/30/is_parler_a_freer_alternative_to_twitter_well_for_now_143580.html

[15] https://blog.twitter.com/en_us/topics/company/2018/measuring_healthy_conversation.html

[16] https://www.abc.net.au/news/2021-01-11/twitter-removes-chinese-embassy-tweet-uyghur-women-baby-machines/13046494

[17] https://www.foxbusiness.com/technology/twitter-iran-leader-tweets-defense

[18] https://www.buzzfeednews.com/article/craigsilverman/facebook-failed-stop-the-steal-insurrection

[19] https://www.forbes.com/sites/kalevleetaru/2017/12/29/facebooks-deletion-of-ramzan-kadyrov-and-who-controls-the-web/

[20] https://www.washingtonpost.com/world/asia_pacific/facebook-blocks-accounts-of-myanmars-top-general-other-military-leaders/2018/08/27/da1ff440-a9f6-11e8-9a7d-cd30504ff902_story.html

[21] https://www.realclearpolitics.com/articles/2020/09/25/how_social_media_platforms_are_narrowing_the_first_amendment_144306.html

[22] https://www.realclearpolitics.com/articles/2020/07/16/social_media_censors_government_--_muzzling_democracy_itself_143727.html

[23] https://www.foxnews.com/opinion/twitter-tried-censor-me-they-lost-sen-tom-cotton

[24] https://www.nytimes.com/2019/07/09/us/politics/trump-twitter-first-amendment.html

stations were barred by law from most censorship of federal political candidates, even when they believed them to be presenting egregious falsehoods or threatening violence, out of concern that no private company should be permitted to censor government. [25]

Perhaps more troubling is that speech rules no longer just govern social spaces. Uber, Lyft and Airbnb have all banned their services [26] from being used by those [27] whose online and offline political speech was deemed unacceptable. [28] Facebook last year extended its reach to the offline world, banning certain kinds of calls for protest [29] while permitting others. [30]

It was a remarkable sight in the aftermath of the January 2021 capitol storming to behold Democratic lawmakers [31] and the press [32] lamenting that Congress does not have the power to silence voices with whom it disagrees and instead urging Silicon Valley to exercise the power only it holds: the ability to silence any voice from the digital world. And this plea came from the very lawmakers who had once condemned social platforms as dangerous monopolies. As if to remind Silicon Valley that Congress controls its destiny, Senator Elizabeth Warren publicly warned Amazon two months later that criticism of Congress would no longer be tolerated. [33]

To some, Silicon Valley's newfound emphasis on combating "misinformation," [34] with private companies as curators of permissible speech and definers of "truth," might seem like a positive development. After all, threats of violence, racism, sexism, doxing, sedition, harmful medical advice and the like are damaging to society. Yet billionaires that can silence presidents, a Congress that can silence dissent and private companies deciding what is "best" for the nation and what constitutes "truth" pose an existential threat to democracy. In the end, the very future of our shared society hinges on the ability of Silicon Valley to balance thoughtful moderation with freedom of speech.

Social media is the communications fabric that underlies modern society and undergirds democracy itself. It is the public square through which we have our society debates. It is the medium through which we speak to our elected officials and they speak back to us. It is a growing channel through which governments from local to national publish laws, policies and regulations to the public, how schools announce schedules, how companies announce products. It is where we talk to the world and where we talk to each other. The censorship rules that social platforms devise thus shape our lives and the future of

[25] https://supreme.justia.com/cases/federal/us/360/525/
[26] https://www.nbcnews.com/news/us-news/laura-loomer-banned-uber-lyft-after-anti-muslim-tweetstorm-n816911
[27] https://news.yahoo.com/why-airbnb-canceled-reservation-customer-171902808.html
[28] https://www.washingtonian.com/2021/01/07/airbnb-says-it-canceled-some-hate-group-members-dc-reservations-and-plans-to-do-more/
[29] https://www.realclearpolitics.com/articles/2020/04/22/facebooks_covid-protest_ban_renews_censorship_concerns_143003.html
[30] https://www.realclearpolitics.com/articles/2020/10/14/social_medias_role_in_democracy_more_harmful_than_helpful_144436.html
[31] https://www.reuters.com/article/us-usa-election-trump-social/senior-u-s-democrat-urges-twitter-facebook-to-ban-trump-from-platforms-idUSKBN29C022
[32] https://www.zdnet.com/article/twitter-should-immediately-and-permanently-ban-trump/
[33] https://twitter.com/SenWarren/status/1375283617341968385
[34] https://www.reuters.com/article/us-usa-election-socialmedia-eu/us-capitol-siege-heralds-tougher-social-media-curbs-says-eu-commissioner-idUSKBN29G11S

our nation in unprecedented ways. The rules they devise become, in many ways, the rules of our national conversations about the future of America.

Does America Still Believe In Free Speech?

Does American society still believe in its founders' vision of freedom of expression as the bedrock of democracy?

In 1939, when Gallup asked Americans "do you believe in freedom of speech," 96% of respondents said yes. When asked whether "radicals" should be granted those freedoms, just 40% of the public agreed, dropping to 36% for communists. [35] Even a century ago there was a stark conflict between the ideal of free speech and the reality of societies wishing to silence disagreeable speech.

Today, 40% of millennials believe that the government should outlaw "statements that are offensive to minority groups," [36] while 78% of college students believe racial slurs should be banned on campuses. [37] More than a quarter of students believe that campuses should ban certain political views. [38] Yet, just as their successors a century ago, 80% of college students agree that even "offensive or biased" speech should be permitted. [39] [40] When the abstract ideal of free speech meets the reality of the kind of expression it permits, Americans today and those a century ago seem to agree there should be limits.

At the same time, the right to free speech is becoming increasingly partisan. In 2017, 47% of Republicans and 44% of Democrats believed that the right of Americans to "be able to speak their minds freely online" was more important than for everyone to "feel welcome and safe online." [41] In just three years the two parties have grown sharply apart, to 60% of Republicans and 45% of Democrats believing in being able to speak openly on the web. [42]

When it comes to combatting falsehoods online, 39% of the public believes the government should be empowered to remove "false information online, even if it limits freedom of information," while 56% of the public believes technology companies should take on this role. As with speech itself, there is a sharp partisan split. Roughly equal percentages of Republications (37%) and Democrats (40%) trust the

[35] https://news.gallup.com/vault/206465/gallup-vault-tolerance-free-speech-limits.aspx
[36] https://www.pewresearch.org/fact-tank/2015/11/20/40-of-millennials-ok-with-limiting-speech-offensive-to-minorities/
[37] https://www.insidehighered.com/news/2020/05/05/gallupknight-foundation-survey-shows-students-conflicted-about-free-speech
[38] https://www.insidehighered.com/news/2020/05/05/gallupknight-foundation-survey-shows-students-conflicted-about-free-speech
[39] https://www.insidehighered.com/news/2020/05/05/gallupknight-foundation-survey-shows-students-conflicted-about-free-speech
[40] https://knightfoundation.org/reports/the-first-amendment-on-campus-2020-report-college-students-views-of-free-expression/
[41] https://www.pewresearch.org/fact-tank/2017/07/24/democrats-more-likely-than-republicans-to-say-online-harassment-is-a-major-problem/
[42] https://www.pewresearch.org/fact-tank/2020/10/08/partisans-in-the-u-s-increasingly-divided-on-whether-offensive-content-online-is-taken-seriously-enough/

government to combat online falsehoods, but when it comes to trusting technology companies, 60% of Democrats and just 48% of Republicans trust them. [43]

History has shown that neither solution is a panacea. Those who trust private companies to faithfully combat falsehoods would do well to reflect on radio's early history of silencing candidates and topics with whom it disagreed, [44] while those who trust government might wish to learn more about how presidential administrations through the years leveraged their power over broadcasters to silence criticism. [45] China shows how dangerous governmental oversight in particular can be, as it leveraged similar powers in April 2021 to silence a number of women's rights groups after deeming them "illegal or hurtful" to society. [46]

Remarkably, even the American Civil Liberties Union has adjusted its once-absolute stance on freedom of expression. The ACLU once routinely defended even the Ku Klux Klan [47] [48] [49] and in 2017 represented the organizer of the "Unite the Right" rally in Charlottesville, Virginia, successfully overturning the city's efforts to move the rally. [50]

In the rally's aftermath, one of the ACLU's Virginia board members resigned, arguing that "what's legal and what's right are sometimes different." [51] The organization itself issued new guidelines [52] that clarified "Our defense of speech may have a greater or lesser harmful impact on the equality and justice work to which we are also committed, depending on factors such as the (present and historical) context of the proposed speech; the potential effect on marginalized communities; the extent to which the speech may assist in advancing the goals of white supremacists or others whose views are contrary to our values; and the structural and power inequalities in the community in which the speech will occur." [53] It subsequently clarified that it would still represent white supremacists in "appropriate circumstances." [54]

When it comes to the freedom of the press, even the Supreme Court has taken a steadily less positive view. As one recent study put it, "A generation ago, the court actively taught the public that the press was a check on government, a trustworthy source of accurate coverage, an entity to be specially protected from regulation and an institution with specific constitutional freedoms … Today, in contrast, it almost never speaks of the press, press freedom or press functions, and when it does, it is in an overwhelmingly less positive manner." [55] [56] In 2019, Justice Clarence Thomas went so far as to argue that the media today

[43] https://www.journalism.org/2018/04/19/americans-favor-protecting-information-freedoms-over-government-steps-to-restrict-false-news-online/

[44] https://repository.uchastings.edu/cgi/viewcontent.cgi?article=1790&context=hastings_comm_ent_law_journal#page=9

[45] https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history

[46] https://www.wsj.com/articles/china-stresses-family-values-as-more-women-put-off-marriage-childbirth-11618824601

[47] https://www.aclu.org/press-releases/aclu-praises-cleveland-mayors-support-kkks-first-amendment-right-march

[48] https://www.aclu.org/press-releases/aclu-em-defends-kkks-right-free-speech

[49] https://www.aclu.org/blog/free-speech/equality-justice-and-first-amendment

[50] https://www.cbsnews.com/news/why-the-aclu-defends-white-nationalist-free-speech-60-minutes/

[51] https://twitter.com/waldojaquith/status/896566113974317058

[52] https://www.cbsnews.com/news/why-the-aclu-defends-white-nationalist-free-speech-60-minutes/

[53] https://www.aclu.org/sites/default/files/field_document/aclu_case_selection_guidelines.pdf

[54] https://www.wsj.com/articles/aclu-isnt-backing-away-from-free-speech-1530024182

[55] https://dnyuz.com/2021/04/19/the-supreme-courts-increasingly-dim-view-of-the-news-media/

[56] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787709

endeavor to “titillate rather than to educate and inform” [57] and thus are undeserving of some of the protections specifically afforded them. [58]

Even the press itself increasingly believes in a narrowing of the First Amendment’s protection, including for their own industry. CNN refers to Fox News as a “propaganda platform” [59] and openly calls for it to be silenced. [60] NBC’s Lester Holt argues that "I think it's become clearer that fairness is overrated … The idea that we should always give two sides equal weight and merit does not reflect the world we find ourselves in. That the sun sets in the west is a fact. Any contrary view does not deserve our time or attention." [61]

At the same time, as the rich diversity of the world’s underrepresented voices are increasingly heard, how are the inevitable conflicts managed in a shrinking landscape of acceptable speech? Should music by one underrepresented community be deleted for attacking another underrepresented community? [62] Which underrepresented communities are permitted to criticize which other underrepresented communities? [63] [64] Which underrepresented communities should be protected from caricature and which should receive no protections from calls for their extermination? [65] [66]

Can The Web Exist Without Moderation?

The great promise of social media was, like all Silicon Valley dreams, to “disrupt” [67] the status quo. For social platforms, that meant reimagining the role of communications and publication in the public sphere, to give voice to the voiceless by creating an unmoderated free speech zone in which all voices were equal. It was seen as “a powerful tool for equalizing imbalances of power by giving voice to the disenfranchised and by allowing more democratic participation in public discourse.” [68]

Yet from its earliest days the web confronted the reality that when diverse societies with very different beliefs, backgrounds and lived experiences come together, there will inevitably be disagreement.

The Usenet newsgroups that formed the “social media” of the early web were filled with “innuendo, sarcasm, obscenities and violent personal abuse and vilification” to the point that “perhaps the most striking communication phenomenon on the Usenet is the frequency of highly uninhibited expression … easily provoked verbal aggression, and disclosure of very personal information.” [69] The concept of “flaming” another user with abuse was commonplace to the point there was even a dedicated newsgroup

[57] https://www.law.cornell.edu/supremecourt/text/449/560
[58] https://www.nytimes.com/2019/02/19/us/politics/clarence-thomas-first-amendment-libel.html
[59] https://www.cnn.com/2021/03/17/politics/ron-desantis-covid-florida/index.html
[60] https://www.wsj.com/articles/just-asking-for-censorship-11614295623
[61] https://thehill.com/homenews/media/545803-lester-holt-warns-media-against-giving-a-platform-for-misinformation
[62] https://www.dailymail.co.uk/news/article-9418825/YouTube-employees-slam-bosses-not-taking-rapper-YGs-anti-Asian-Meet-Flockers-video.html
[63] https://deadspin.com/no-one-seems-to-care-that-kevin-durant-is-a-homophobe-1846597309
[64] https://www.theguardian.com/media/2021/mar/14/teen-vogue-alexi-mccammond-tweets-controversy
[65] https://www.dailymail.co.uk/news/article-9326645/Ebay-BANS-people-reselling-six-offensive-Dr-Seuss.html
[66] https://www.newsweek.com/ebay-removes-discontinued-dr-seuss-books-1573824
[67] https://www.theguardian.com/news/2020/sep/24/disruption-big-tech-buzzword-silicon-valley-power
[68] https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1071&context=dlj
[69] https://files.eric.ed.gov/fulltext/ED334620.pdf

for attacking others that many users considered their "home." [70] Abuse was so severe that a few groups even split off so-called "nice" editions in which such attacks were banned. Foreshadowing today's controversial debates, "newsgroups with the largest readership and heaviest traffic (e.g. talk.politics.mideast and talk.politics.misc) have an extremely confrontational flavor, with argument frequently escalating into personal abuse, obscenities, and vilification." [71]

The debate over whether such speech should be silenced was already raging even in this early era. In January 1989, Stanford University announced it would be blocking access to a Usenet humor newsgroup over its publication of racist and sexist jokes. Faculty in the computer science department opposed the proposed ban as an infringement on freedom of speech and censorship and launched an online petition to overturn it, arguing that "In a nation in which many of the major media organizations are controlled by large corporations, the existence of anarchical news groups are refreshing and necessary in a democracy." [72]

In contrast, an African-American student in the department supported the ban, offering that "even if I can't force the presentation of other cultures - and I DO NOT assume this is impossible - I will ALWAYS protest the stereotyping of my culture … whether disguised as free speech or simply stated as racism or sexism, such humor IS hurtful. It is a University's right and RESPONSIBILITY to minimize such inflammatory correspondence in PUBLIC telecommunications." [73] A decade and a half before Facebook's founding, the tension between free speech and hateful speech was raging.

In justifying its ban, the University offered that "there mere reason that these jokes are offensive is not enough to shut the system down. This particular news group does not provide a mechanism in which it can be discussed." [74] (The humor group did not permit users to post or comment themselves). In other words, in the University's view, the offensive speech was permissible so long as users could respond to it to debate and condemn it.

Moderation In Action: Wikipedia

What does content moderation look like in practice? Much has been written [75] on what social platform content moderation really looks like behind closed doors. The endless streams of horrors and hate [76] reviewers must confront each day reminds us just how toxic the online world can be. Yet while conveying the dark depths of the online world, these documentaries detail a world of contested and complex moderation decisions that can seem abstract to the typical internet user. After all, many of the horrific stories of child abuse, livestreamed tortures and murders can seem far away from the everyday concerns of hateful and threatening speech experienced by most users.

[70] https://files.eric.ed.gov/fulltext/ED334620.pdf
[71] https://files.eric.ed.gov/fulltext/ED334620.pdf
[72] https://archives.stanforddaily.com/1989/01/30?page=1&section=MODSMD_ARTICLE9#article
[73] https://www.newyorker.com/tech/annals-of-technology/origin-silicon-valley-dysfunctional-attitude-toward-hate-speech
[74] https://archives.stanforddaily.com/1989/01/30?page=1&section=MODSMD_ARTICLE9#article
[75] https://yalebooks.yale.edu/book/9780300235883/behind-screen
[76] https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity

Without actually serving as a content moderator, it can be hard for an ordinary internet user to understand why content moderation is so complex. Social media platforms are careful to perform their moderation entirely outside public view, meaning there are few opportunities for the public to see firsthand just how controversial and contested moderation decisions can be.

Wikipedia confronts a microcosm of these decisions each day as its contributors and administrators debate everything from what warrants inclusion to how it is presented and what evidence is cited. Its coverage of almost every topic imaginable and realtime updating means these debates play out every day on topics from the most mundane updates to the most controversial breaking stories.

Unlike most other moderated platforms, however, Wikipedia makes this debate transparent, with the "Talk" page of any entry offering a live chronology of these debates. Browsing these Talk pages, especially those governing controversial topics of the moment, offers a glimpse into just how caustic, personal and adversarial these debates can become, filled with name-calling, personal attacks, screaming matches and administrator intervention. From minor disputes over a citation or reference, these debates can rage over far more existential questions, such as whether the subject of a page "is a prominent and influential scientist with wide community support" or a "fringe pseudoscientist that claims to have conversed with aliens." [77] The resulting decisions determine what constitutes "truth" to the automated gatekeepers that manage the digital world, from what we see in our web searches to what our smart speakers and phones answer to our questions.

The way in which Wikipedia presented the sexual assault allegations against Joe Biden and Brett Kavanaugh captures the powerful influence of these debates. Nearly a third of the opening text of Kavanaugh's entry details the sexual assault allegations against him, while much of the debate on the Talk page for his entry centers on what sources to cite and word choices, rather than whether those allegations should be mentioned.

In contrast, the allegations against Joe Biden received just a single mention near the bottom of his entry for much of the first half of 2020, [78] with three sentences describing them and three denying them, one from the Biden campaign and two from a New York Times article. Discussion on the entry's "Talk" page emphasized whether the allegations should be mentioned at all and whether they should be seen as credible. [79]

Only later were the allegations against Biden expanded and given their own page. [80] The Talk page for the entry shows just how divisive and controversial the editors found mentioning the accusations at all. [81] In fact, at one point a group of editors voted 18 to delete the entry and 37 to keep, but since there was no consensus under Wikipedia rules, an administrator left the page intact. [82] Had a dozen votes changed out of this small group, the allegations against Biden might have largely disappeared from Wikipedia. In the #MeToo era it is remarkable that a small group of anonymous individuals held a vote to decide whether

[77] https://en.wikipedia.org/wiki/Talk:Jean-Pierre_Petit
[78] https://en.wikipedia.org/w/index.php?title=Joe_Biden&oldid=950966184
[79] https://www.realclearpolitics.com/articles/2020/04/29/biden_vs_kavanaugh_how_the_metoo_numbers_stack_up_143065.html
[80] https://en.wikipedia.org/wiki/Joe_Biden_sexual_assault_allegation
[81] https://en.wikipedia.org/wiki/Talk:Joe_Biden_sexual_assault_allegation
[82] https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/Log/2020_April_12#Joe_Biden_assault_allegation

Biden's accusers were credible and whether their stories should be believed or even heard and that their decision holds so much weight in a digital world that relies so heavily on Wikipedia.

Wikipedia is often held up as an "unbiased" and "neutral" resource in which the world's citizens come together to "debat[e] [their] way to consensus." [83] Moderation is largely focused on determining what persons and topics are "notable" enough to warrant their own entries, what sources are "reputable" enough to cite and what statements are "factual" enough to cite. Yet despite being the knowledge source for our smart speakers, our search engines and social platforms and being lauded by pundits as "unbiased," [84] the biographies, backgrounds and demographics of Wikipedia's contributors and administrators are not always easily accessible or obviously relevant to the topics they edit.

In a traditional encyclopedia, subject matter experts with deep expertise in each topic are recruited to write and edit each piece. On Wikipedia, no such qualifications are required. In 2019 the Washington Post profiled a 36-year-old academic physicist who in his spare time was helping to edit the entry on Hunter Biden. After seeing Biden's entry fill up with references to his business dealings in Ukraine, the physicist "had to get in there and clean it out like a garbage disposal," [85] replacing what he saw as pro-Trump narratives and citations with those of outlets like PolitiFact, Bloomberg and the Washington Post. Other users deleted and restored references to Biden's relationship with his late brother's widow, arguing over whether such information was relevant to the public debate. However, unlike on social media platforms, all of this debate and editing is recorded for posterity, allowing the public to see just how contentious these debates can be.

At the same time, the demographics of who contribute to Wikipedia have historically hardly been representative of society at large. The site's majority male editors [86] have over the years led to a site that has minimized the role of women in STEM fields. [87] Moreover, as efforts were launched to better represent women scientists on Wikipedia, some editors moved swiftly to delete entries or lump them under their husbands, [88] [89] arguing that many women scientists weren't noteworthy enough to warrant their own Wikipedia entries. Similar concerns have been raised about its representation and coverage of other underrepresented groups like racial minorities. [90] [91]

It is important to recognize this critical distinction between transparency and bias. A platform can be highly transparent but at the same time have significant biases, as Wikipedia's co-founder Larry Sanger

[83] https://www.cjr.org/special_report/building-honest-internet-public-interest.php
[84] https://www.cjr.org/special_report/building-honest-internet-public-interest.php
[85] https://www.washingtonpost.com/politics/checking-the-web-on-hunter-biden-a-36-year-old-physicist-helps-decide-what-youll-see/2019/09/25/16573a1e-df9c-11e9-be96-6adb81821e90_story.html
[86] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia
[87] https://www.nature.com/articles/d41586-018-05947-8
[88] https://www.chemistryworld.com/news/female-scientists-pages-keep-disappearing-from-wikipedia-whats-going-on/3010664.article
[89] https://undark.org/2019/04/25/wikipedia-diversity-problem/
[90] https://www.chicagotribune.com/entertainment/ct-ent-black-artists-wikipedia-0927-20200924-moomhhdy3bcthnmyit6go2x23e-story.html
[91] https://slate.com/technology/2020/06/wikipedia-george-floyd-neutrality.html

argued [92] [93] earlier this year. Simply because Wikipedia's rules, debates and edits are public does not free them from the potential of bias, but that transparency does allow researchers and journalists to document [94] these trends and open them to public debate. Indeed, the public debate and rule changes that have followed past leaks of internal social platform moderation guidelines and research offers a preview of the impact that true transparency around social platform moderation could have on the invisible hands that increasingly shape our national democratic debates.

The backlash from some Wikipedia editors to the creation of new entries for female scientists reminds us of the dangers of demographically skewed content moderators. Yet both Twitter and Facebook have historically refused to release detailed demographic breakdowns of their moderators and any biases observed in their actions, making it impossible to know whether the platforms suffer similar unconscious biases. [95] [96]

What sources are citable on Wikipedia? Fox News has been deemed by Wikipedia's editors as an unreliable source for many topics and thus can't be cited, [97] while MSNBC is in Wikipedia's eyes a reliable, neutral and trustworthy source for all topics, including politics. [98] Similarly, "there is consensus that the New York Post is generally unreliable for factual reporting" while "a 2020 RfC found HuffPost staff writers fairly reliable for factual reporting on non-political topics." [99] Once again, Wikipedia's transparency means these decisions are available for public debate, whereas the internal lists used by social platforms are highly secretive. [100] [101]

Wikipedia allows us to see in microcosm the complexities that surround moderation. At the same time, it represents a best-case scenario in which large teams of editors are able to converse, debate, research and evolve their decisions over time. In contrast, on social media, moderators must make decisions by themselves in seconds without the benefit of context, time to conduct additional research or consult external experts.

In short, through Wikipedia we can see just how difficult moderation really is and how even simple questions can elude consensus.

Moderation Is Everywhere

[92] https://nypost.com/2021/07/16/wikipedia-co-founder-says-site-is-now-propaganda-for-left-leaning-establishment/
[93] https://www.realclearpolitics.com/video/2021/07/22/wikipedia_co-founder_on_bias_website_has_completely_abandoned_neutral_point_of_view_sullied_reputations.html
[94] https://theintercept.com/2020/07/02/kamala-harris-wikipedia/
[95] https://www.forbes.com/sites/kalevleetaru/2018/01/12/is-twitter-really-censoring-free-speech/
[96] https://www.forbes.com/sites/kalevleetaru/2016/05/12/does-facebook-suffer-from-unconscious-bias-an-insider-view-into-human-cataloging/
[97] https://www.realclearpolitics.com/articles/2020/08/07/social_media_imposing_modern-day_hays_code_on_political_speech_143911.html
[98] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources
[99] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources
[100] https://www.forbes.com/sites/kalevleetaru/2016/05/09/is-facebook-censoring-conservative-news-how-social-media-controls-what-we-see/
[101] https://www.forbes.com/sites/kalevleetaru/2016/05/13/is-facebooks-trending-topics-biased-against-africa-and-the-middle-east/

Any effort to understand social platform speech policies must look at them as an intrinsic component of our emerging digital world. The concept of content moderation and censorship on the web is typically associated today with user-generated content platforms like social media. Yet acceptable speech policies are increasingly spreading across the online and offline worlds, from ride sharing and room renting to cloud vendors and desktop software. We have to approach the question of content moderation not through the narrow lens of social media companies, but rather through the fact that an ever-growing fraction of our modern lives is now subject to acceptable speech policies established by private companies.

For the first time in American history, government is subordinate to private censorship. Rather than establish policy or enforce compliance, as it has with every other communications technology since the nation's founding, even the government itself is now censored by private individuals, as President Trump discovered.

The growing monopolies of the digital era and the vast archives of personal data they compile on us is enabling an entirely new world of whole-of-life censorship. In the place of China's government-controlled "Social Credit System," [102] private companies are increasingly reaching across our digital lives and collaborating together to banish those they dislike.

In November 2017, Uber and Lyft banned Laura Loomer over anti-Muslim tweets. [103] Three months earlier, Airbnb mass-canceled reservations for those attending events related to the Unite The Right rally in Charlottesville and announced it would preemptively cancel all reservations relating to similar events, stating that "In 2016 we established the Airbnb Community Commitment [requiring users to] accept people regardless of their race, religion, national origin, ethnicity, disability, sex, gender identity, sexual orientation, or age." [104] In November 2020, after an activist Twitter account dedicated to "research and analysis of the far right" provided Airbnb with evidence that a Proud Boys member had reserved an Airbnb property for the Million MAGA March, the platform banned him [105] and reiterated that all hate group members were barred from its service. [106]

In January 2021 the company canceled all reservations in the DC area ahead of the inauguration, citing "reports emerging yesterday afternoon regarding armed militias and known hate groups that are attempting to travel and disrupt the Inauguration." [107] The company further noted that "on an ongoing basis, Airbnb has removed people from the platform associated with violent hate groups in advance of specific events" and that "Members of hate groups are never welcome on Airbnb and we have previously taken action to remove these individuals from the Airbnb community." [108]

Yet in the absence of federal law banning hate speech, there is no settled legal precedent for deciding just what counts as "hate speech" or "hate group" membership. Is membership in an underrepresented community organization that calls for the extermination of another underrepresented community a "hate

[102] https://en.wikipedia.org/wiki/Social_Credit_System
[103] https://www.nbcnews.com/news/us-news/laura-loomer-banned-uber-lyft-after-anti-muslim-tweetstorm-n816911
[104] https://gizmodo.com/airbnb-won-t-put-a-roof-over-the-heads-of-nazis-1797585928
[105] https://news.yahoo.com/why-airbnb-canceled-reservation-customer-171902808.html
[106] https://www.fastcompany.com/90575291/airbnb-is-canceling-bookings-by-members-of-hate-groups-headed-to-the-million-maga-march
[107] https://news.airbnb.com/airbnb-to-block-and-cancel-d-c-reservations-during-inauguration/
[108] https://news.airbnb.com/airbnb-announces-capitol-safety-plan-for-the-inauguration/

group" if it itself has been discriminated against? In the absence of the kinds of bodies of case law that exist in countries with criminal hate speech prohibitions, [109] private companies are left to make their own decisions.

Social platforms are also increasingly extending their reach to the offline world. In April 2020, Facebook announced it would begin banning real-world protest events organized on its platform that "defy government's guidance on social distancing," specifically citing anti-lockdown protests. [110] At the same time, the company quietly exempted certain other protests against the advice of government health authorities. [111]

As access to movies, music, books, scientific and medical literature is increasingly centralized in a handful of streaming services, a ban from even a few of these services can effectively sever our access to an ever-growing fraction of the world's knowledge and entertainment, while their decisions about what is acceptable can shape the public debate. Amazon, which accounts for 72% of new adult book sales, [112] increasingly removes books with which it disagrees. [113] eBay now removes Dr. Seuss books containing racist caricatures but permits the sale of Mein Kampf and other works advocating violence against those of Jewish descent. [114] [115]

Even the most basic economic building blocks of society now come with speech rules attached. Payments processing companies now ban political organizations they disagree with [116] and banks leverage their power over the financial system to exclude disagreeable speech. [117]

Nearly every software product, service, app, website today includes some kind of end user agreement that gives the company the right to determine at its sole discretion whether the person's speech is unacceptable and ban them. Such rules appear in the most unexpected places. Since at least 2016, [118] Microsoft has included provisions in its terms of service banning the use of its Office 365 software to "engage in activity that is harmful to you, the Services or others," including "communicating hate speech." Microsoft "reserves the right to review Your Content" at its sole discretion and if it determines that a user's speech violates these rules, it can ban the user from Office 365 and even the Microsoft Account they use to log into their desktops, laptops and tablet devices. Asked in 2019 how it defined "harmful" or

[109] https://en.wikipedia.org/wiki/Hate_speech_laws_in_Canada
[110] https://www.realclearpolitics.com/articles/2020/04/22/facebooks_covid-protest_ban_renews_censorship_concerns_143003.html
[111] https://www.realclearpolitics.com/articles/2020/10/14/social_medias_role_in_democracy_more_harmful_than_helpful_144436.html
[112] https://www.wsj.com/articles/they-own-the-system-amazon-rewrites-book-industry-by-turning-into-a-publisher-11547655267
[113] https://abigailshrier.substack.com/p/book-banning-in-an-age-of-amazon
[114] https://www.dailymail.co.uk/news/article-9326645/Ebay-BANS-people-reselling-six-offensive-Dr-Seuss.html
[115] https://www.newsweek.com/ebay-removes-discontinued-dr-seuss-books-1573824
[116] https://www.wsj.com/articles/stripes-ban-on-trump-campaign-isnt-absolute-11610632728
[117] https://www.realclearinvestigations.com/articles/2021/01/10/its_back_the_political_struggle_for_control_of_banks_loan_taps_126710.html
[118] https://web.archive.org/web/20160728121715/https:/www.microsoft.com/en-us/servicesagreement/upcoming.aspx

"hate" speech and whether it had banned anyone for such speech, the company declined to comment, offering a reminder of just how opaque these rules are. [119]

The growing consolidation of the data centers that power the internet and the mobile devices we use to access it means the internet itself is increasingly coming with default minimum speech rules. In response to what it saw as the growing censorship of Twitter, Parler [120] emerged as a Twitter-like social platform that performed only minimal content moderation, reaching number one on Apple's App Store in January 2021. [121] Yet within days Apple and Google had banned the download of it from their respective app stores, [122] effectively banishing it from mobile devices across the country. [123] Parler's cloud hosting provider, Amazon Web Services, evicted it, [124] taking the site offline until a conservative cloud provider agreed to host it. [125] Thus, any website, app or other digital service must comply with the censorship policies of Silicon Valley or it will be denied access to the plumbing of the internet.

Even offline media are not immune. Television channels must contract with cable carriers to transmit them into homes, syndicated radio shows must be hosted by stations, and even independent newspapers must have websites and mobile apps. With local news outlets diminishing, it is important to note that no matter how editorially independent some may be, all are still dependent on cloud providers, app stores, and Internet service providers.

As technology makes it possible for companies to reach ever-deeper into our lives, advances like realtime voice censorship [126] mean it is only a matter of time before the phone company begins to censor our private conversations in realtime. As our homes are increasingly filled with smart appliances, soon our internet-connected toasters and smart thermostats will come with acceptable speech rules and will turn themselves off at the first errant tweet.

In short, the content moderation of social media platforms now increasingly extends to every facet of our lives, online and offline.

The Logical Conclusion: Facebook's "Supreme Court"

Who decides the rules of the web?

Left to unmoderated anarchy, the history of the web's early social media platforms like Usenet reminds us how easily platforms can devolve into cesspools of threats of violence, doxing, harassment, personal demographic attacks, blatantly harmful medical or financial information and illegal content. This raises the question of how to permit society to express itself as freely as possible without devolving into harm?

[119] https://www.forbes.com/sites/kalevleetaru/2019/07/26/censorship-comes-for-the-desktop-how-microsoft-has-infused-values-into-windows-and-office/
[120] https://en.wikipedia.org/wiki/Parler
[121] https://www.cnn.com/2020/12/10/tech/parler-downloads/index.html
[122] https://www.theverge.com/2021/1/9/22221730/apple-removes-suspends-bans-parler-app-store
[123] https://variety.com/2021/digital/news/parler-app-banned-google-apple-app-store-trump-violence-1234881947/
[124] https://www.buzzfeednews.com/article/johnpaczkowski/amazon-parler-aws
[125] https://www.bbc.com/news/technology-55615214
[126] https://www.wsj.com/articles/dont-think-just-bleep-11618162999

Social platforms today employ vast armies of human reviewers [127] and an ever-expanding landscape of opaque computer algorithms to police their walled gardens. [128] In a diverse society there will be disagreements, especially when the moderators reviewing content may be contractors halfway across the world given just seconds to review content that taps deeply into lived experiences and local culture with which they may be entirely unfamiliar. Many decisions may not have an easy "right" answer with little consensus no matter how many reviewers are consulted.

Much as the United States' court system has a "Supreme Court" that can resolve such disputes, Facebook announced in November 2018 its "Oversight Board" which has effectively similar powers over the company's content moderation decisions, announcing its initial members in May 2020. [129] Any moderation decision on Facebook or Instagram can be appealed to the Board which will select cases that are "difficult, significant and globally relevant that can inform future policy." [130]

In addition to appealing removals of their own content, as of April 2021, users can also appeal Facebook's refusal to remove someone else's content. A user that reports a piece of content to Facebook as something they object to and which Facebook declines to remove, can now appeal to the Board to silence speech from afar. [131] The Board also accepts commentary from the general public in weighing its decisions. [132]

Unlike the US Supreme Court, which is appointed by elected officials, the citizenry of the world has no say over Facebook's Board and cannot influence it if they feel it does not reflect societal consensus. Its global jurisdiction all but guarantees that in a diverse world its decisions will reflect the natural tension between the speech rights of one group and the human rights of another.

For example, three years ago the United Nations singled out Facebook's role [133] in spreading hate speech against Muslims in Burma that helped fuel genocide.[134] Yet in January 2021 the Oversight Board ruled that despite Facebook's concerns over "anti-Muslim hate speech," statements by Burmese users like "something's wrong with Muslims psychologically" or "male Muslims have something wrong in their mindset" do not violate [135] Facebook's Hate Speech rules that prohibit "dehumanizing speech, harmful stereotypes, statements of inferiority" and "generalizations that state inferiority." [136] While conceding that "hate speech against Muslim minority groups in Myanmar is common and sometimes severe," it was the Board's conclusion that "while the terms used could show intolerance, they were not derogatory or violent." [137]

---

[127] https://www.facebook.com/journalismproject/facebook-oversight-board-for-content-decisions-overview

[128] https://www.realclearpolitics.com/articles/2020/07/12/facebook_audit_exposes_algorithm_biases_in_policing_speech.html

[129] https://en.wikipedia.org/wiki/Oversight_Board_(Facebook)

[130] https://oversightboard.com/appeals-process/

[131] https://about.fb.com/news/2021/04/users-can-now-appeal-content-left-up-on-facebook-or-instagram-to-the-oversight-board/

[132] https://www.bbc.com/news/technology-56781104

[133] https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN

[134] https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

[135] https://www.oversightboard.com/decision/FB-I2T6526K

[136] https://www.facebook.com/communitystandards/hate_speech/

[137] https://www.oversightboard.com/decision/FB-I2T6526K

The Board went on to conclude: [138]

> *The Board acknowledges that online hate speech in Myanmar has been linked to serious offline harm, including accusations of potential crimes against humanity and genocide. As such, the Board recognized the importance of protecting the rights of those who may be subject to discrimination and violence, and who may even be at risk of atrocities. Nonetheless, the Board concludes while some may consider the post offensive and insulting towards Muslims, the Board does not consider its removal to be necessary to protect the rights of others.*

The Board noted that its decision rested in part on the user's claim that they were merely condemning religious extremism and that "the fact that the post was within a group that claimed to be for intellectual and philosophical discussion, and also drew attention to discrimination against Uyghur Muslims in China, lends support to the user's claim." Under the Board's rationale, would replacing "Muslim" with "African American" be acceptable so long as the post occurred in an "intellectual and philosophical" debate board that also drew attention to racism?

At what point does a post attacking Muslims cross the line from apparently acceptable mere "intolerance" to prohibited "derogatory" speech? The world's 1.8 billion Muslims [139] have no say in this debate. Despite having authority over Facebook in every country it is used, the Board's members have "lived in" just 27 countries and speak just 29 languages. [140] This despite representing a world of at least 193 countries [141] and more than 7,000 languages [142] (though some countries ban access to its services and not all languages may be supported by its tools). Moreover, while members of the US Supreme Court are appointed and confirmed by democratically elected officials representing the citizenry of the nation, the world's Facebook users have no say in the membership or composition of Facebook's Board. If the citizens of a nation believe they are not represented by Facebook's Board and that its decisions are empowering genocide against them, they have no recourse to force change.

Perhaps most extraordinary of all, however, is the Board's oversight of the world's elected governments, including former presidents. [143] In March 2021, Facebook clarified that the Board has oversight over its "elected officials policy … from the President of the United States to your local school board official." [144] As to whether all democratically elected officials or candidates for office should have a voice on social media, the Board's spokesperson offered that there is "a very good argument to be made that an elected official has other ways to communicate without using social media. They have a blog, they have press statements; the press covers them. We will be wading into these very complex issues." [145] As Donald

[138] https://www.oversightboard.com/decision/FB-I2T6526K
[139] https://www.pewresearch.org/fact-tank/2019/04/01/the-countries-with-the-10-largest-christian-populations-and-the-10-largest-muslim-populations/#:~:text=Overall%2C%20there%20are%20about%202.3,world%20and%201.8%20billion%20Muslims.
[140] https://about.fb.com/news/2020/05/welcoming-the-oversight-board/
[141] https://www.un.org/en/about-us
[142] https://www.ethnologue.com/guides/how-many-languages
[143] https://www.bbc.com/news/technology-56781104
[144] https://www.engadget.com/facebook-oversight-board-rules-for-politicians-donald-trump-211353831.html
[145] https://www.engadget.com/facebook-oversight-board-rules-for-politicians-donald-trump-211353831.html

Trump's social ban reminds us, [146] no other communications technology comes close to the visibility of social media today.

Put simply, a group of unelected individuals, including former heads of state of foreign nations, now has the power of censorship over the domestic political debates of the United States' citizenry and the official speech of its democratically elected government on one of the most powerful platforms in the world. With three quarters of its board hailing from outside the United States, the views of the board do not necessarily reflect the views of the American electorate, yet it has the power to silence dissenting views.

While many might cheer the idea of subjecting America's political speech to an international standard more narrow than its own First Amendment, what happens when the views of America differ from the international community? For example, what if the board rules that stacking the US Supreme Court is inconsistent with international standards and bars all discussion by the Democratic Party? In keeping with its ruling on anti-Muslim hate speech, what if it rules that Facebook is barred from removing racism against African Americans so long as it merely exhibits "intolerance?" What if it ruled that Joe Biden must be removed from Facebook ahead of the 2024 election because of allegations of sexual assault against him? At first glance these might appear hyperbolic hypotheticals, but America's long history of censorship experiments reminds us that when it comes to political speech, anything is possible.

Censorship Without Representation

Why do the speech rules of social media companies matter? They matter because the technology to enforce realtime society-scale censorship has arrived without the corresponding societal processes and agreements over what should be censored. For more than 200 years Americans have argued over how to define "acceptable speech" and experimented with almost every form of censorship, to no avail.

The "intractable" problem [147] of defining acceptable speech has eluded all consensus so we have in effect given up as a nation and left it to private companies to sort out on their own. Uniquely in a democracy, the citizenry has no voice under this model, no ability to shape the rules that increasingly govern its speech and no right even to see the rules under which it lives. Social platforms now invisibly shape the speech of democracy, accountable to no-one, with no visibility or transparency and no societal understanding of the impact of their actions on the course of our nation.

We have arrived at a point in history in which the technology exists to censor an ever-growing fraction of human knowledge and communication, while the consolidation of the digital world means a handful of companies now decide the speech of the entire planet. With a few lines of code, a person or idea can simply vanish from the digital world, while AI algorithms are increasingly being turned loose to try and identify the next subversive thought before it can be expressed. We have the power to censor democracies today in a way that even the most repressive regimes of the past could not imagine.

Yet the ease with which we can now censor masks the simple fact that the most important question of all remains unanswered: what to censor.

[146] https://apnews.com/article/donald-trump-television-media-social-media-mar-a-lago-15475effad776332890300bcb2b35bde

[147] https://books.google.com/books?id=9LHQehQel4UC&pg=PR9

In short, we have an era of “censorship without representation” in which private companies are left to decide what is best for the nation.

Seduced by the idea that the precision of mathematics and computer code can solve society’s greatest challenges, we have in effect asked a handful of private companies to solve what two centuries of democracy could not.

To better understand where these trends might take us, it is helpful to understand how we got here and whether the challenges of today’s social platform moderation are truly novel or whether they merely reflect the age-old questions of America’s two-hundred-year experiment with free speech.

A TWO-CENTURY EXPERIMENT IN FREE SPEECH

Almost since its founding, the United States has wrestled with the balance between preserving freedom of expression and the desire by government to constrain the publication of "harmful" speech.

After an early abortive attempt at federal speech rules, the role of censor in America was decentralized and largely left to the cities and states to reflect local sensibilities and debates. As technological advancements improved the distribution of information, this vibrant patchwork of formal and informal censorship brought divergent views into increasing collision and censorship gradually centralized from the states to the federal government to mediate these disagreements. With the rise of motion pictures and eventually broadcasting, private companies took over the role of censor under a system of "mandatory voluntary" rules. Even as government increasingly intruded, day-to-day responsibility for deciding acceptable and disallowed speech fell not to official government censorship officials like in earlier times and in some countries today, but to the companies themselves. Publishers adopted industry-wide guidelines that were officially voluntary, but with government coercion to make them mandatory. Government loomed ever large over these daily censorship decisions and intervened to set precedents and establish minimum speech guarantees, especially regarding political speech, but companies were largely left to devise the details. Over time these cooperative codes gave way to monopoly power with consolidation, with cable television ushering in a gatekeeping model not that dissimilar from today's social platforms, while Section 230 consolidated this power and removed the last remaining state and local influence on speech rules. All the while, the courts guided these societal debates and tried to devise some semblance of consensus over the "intractable" problem of defining societal-wide acceptable speech.

Through this two-century struggle to balance the First Amendment and speech rules, the United States experimented with almost every model of censorship. Early attempts focused on regulating speakers themselves, but over time most efforts refocused on gatekeepers, allowing citizens freedom to express their views under the First Amendment, but limiting the distribution of undesirable views to the public. Early attempts at allowing censorship rules to reflect local concerns gave way to centralized national rules, which social platforms today have today turned into global rules. Allowing states agency to define acceptable speech failed to prevent conflicts, as states attempted to silence speech from afar, while centralizing power meant a single set of rules had to be defined for an entire nation. These speech arbitrators evolved from government officials in the Post Office era to private companies in the motion picture and early radio era to hybrid models in the later broadcasting era. Left in private hands, publishers censored topics and public figures they disliked. Left in government hands, policy dissent and criticism were silenced. Left to the courts, consensus was elusive and the rules ever-changing. In every case, minority voices were silenced. The end result is that to date none of these attempts at regulating speech has yielded a durable consensus that also permitted a wide diversity of voices and perspectives. With the rise of the internet, lawmakers have once again reverted to the privatized censorship model of early broadcasting, this time empowering private companies with near-absolute censorship powers.

While the history of speech regulation and First Amendment battles in the United States has been documented ad infinitum, it is worth reviewing a few highlights from this history that help frame and contextualize the current debates over social media moderation and the 200-year struggle to define "acceptable speech" in the context of America's prioritization of free expression.

Early Censorship & State Control

The fledging nation was just over two decades old and the First Amendment just seven years old, when the Alien and Sedition Acts were passed in 1798. [148] Passed by the Federalists in an attempt to silence Democratic-Republican criticism of the government, [149] the Acts made it a crime to "print, utter, or publish … any false, scandalous, and malicious writing" [150] about the US Government. While officially designed to combat falsehoods, enforcement was largely limited to newspaper editors that criticized the government, setting an early precedent for governmental desire to restrict public speech. The intense public backlash to these laws enshrined the new nation's rejection of formalized federal speech control. It also served as an early warning that attempts to combat falsehoods can be easily abused to silence dissenting speech.

The rise of abolitionist mailings from the North to the South during the 1830s led to a vigilante mob raiding the Charleston, South Carolina post office and burning the abolitionist literature they found. In the raid's aftermath, the US Postmaster General informally permitted southern post offices to refuse to deliver material critical of slavery. [151] [152] The First Amendment's protections meant the Postmaster could not formally censor such content and despite debating the need for federal intervention, Congress ultimately left it to the states to police the mail according to their local laws and morals. Throughout the 1800s and early 1900s censorship at the state and local level was common, leading to a fragmented patchwork of rules in which an idea could be encouraged in one place and banned in another. [153] This permitted censorship laws to reflect the diversity of American communities, while seeding ever-greater clashes as technological advances allowed ideas to transcend geography ever more rapidly and cheaply.

Comstock Laws, Hollywood & National Speech Standards

In 1873, Anthony Comstock [154] of the New York Society for the Suppression of Vice [155] convinced Congress to pass the Comstock Laws [156] that banned using the US Post Office to distribute obscenity, birth control information and sexual content and "every obscene, lewd, or lascivious, and every filthy book, pamphlet, picture, paper, letter, writing, print, or other publication of an indecent character." [157] This helped crystalize the concept of a single national standard of "acceptable speech" rather than leaving it to the states and cities to determine what was acceptable to their respective communities. Centralized speech rules meant there must be a single standard acceptable to the entire population, laying the groundwork for a century of debate over how to create a single set of acceptable speech rules that would satisfy an entire nation.

[148] https://en.wikipedia.org/wiki/Alien_and_Sedition_Acts
[149] https://history.house.gov/Historical-Highlights/1700s/The-Sedition-Act-of-1798/
[150] https://memory.loc.gov/cgi-bin/ampage?collId=llsl&fileName=001/llsl001.db&recNum=719
[151] https://digitalcommons.unomaha.edu/cgi/viewcontent.cgi?article=1478&context=studentwork
[152] https://www.washingtonpost.com/outlook/2020/08/16/politicizing-usps-is-another-andrew-jackson-move-trump/
[153] https://books.google.com/books?id=fYYtAAAAMAAJ&pg=PA232#v=onepage&q&f=false
[154] https://en.wikipedia.org/wiki/Anthony_Comstock
[155] https://en.wikipedia.org/wiki/New_York_Society_for_the_Suppression_of_Vice
[156] https://en.wikipedia.org/wiki/Comstock_laws
[157] https://en.wikipedia.org/wiki/Comstock_laws#Text_of_the_parent_federal_law_for_the_United_States

In 1897 [158] [159] the nascent filmmaking industry encountered its first censorship fights, leading over the next few decades to a flurry of city and state censorship boards to govern [160] the content [161] of this new medium. These early fights foreshadowed the debates that would accompany each new medium to come, from radio to television to the internet.

The Sedition Act of 1918 [162] resurrected many of the goals of its century-old predecessor, barring "disloyal, profane, scurrilous, or abusive language [that brings the government] into contempt, scorn, contumely or disrepute." [163] This occurred alongside the US Government Committee on Public Information [164] that formed a domestic government-supported propaganda machine designed to promote a positive image of the government and wartime efforts. Efforts to extend their reach beyond the end of the war failed, once again reinforcing public concern over formal government control of speech, but also entrenching the idea that speech must be constrained in the interest of national unity and cohesion.

The Hayes Code & Hollywood's "Mandatory Voluntary" Rules

Within a decade, the Motion Picture Producers and Distributors of America released its voluntary "Don'ts and Be Careful's" list in 1927, [165] governing acceptable speech and depictions in movies. This led in 1934 to the Motion Picture Production Code, also known as the Hayes Code. [166] While not enforced by government mandate, the voluntary Code was designed to head off the growing threat of government intervention, with the ever-present threat of the alternative of government-enforced speech rules ensuring compliance.

In many ways Hollywood's "mandatory voluntary" model established the "community standards" models of social platforms today in that Hollywood was left to police itself, but with the specter of government control lurking always in the background to enforce compliance. The primary difference between Hayes-era Hollywood and social media companies today is that local censorship boards across the US had final authority over the studios, while today Section 230 places social media companies beyond the reach of the states and even beyond most federal reach, as evidenced by the inability even of the president of the United States to restore his suspended social media accounts in his final days in office.

The Rise Of Broadcasting & The Normalization Of Government Speech Rules

At the same time, radio was confronting its own crisis of public criticism. Broadcasters faced the looming threat of government regulation of broadcast speech or even nationalization. [167] Concern over the ability to radio to sway the electoral system led in part to the Radio Act of 1927 establishing the "equal time rule" for political advertising and noted that stations "shall have no power of censorship" over that

[158] https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1348&context=yjlh
[159] https://libres.uncg.edu/ir/uncw/f/davisw2008-1.pdf
[160] https://libres.uncg.edu/ir/uncw/f/davisw2008-1.pdf
[161] https://kb.osu.edu/bitstream/handle/1811/67710/1/OSLJ_V13N3_0350.pdf
[162] https://en.wikipedia.org/wiki/Sedition_Act_of_1918
[163] https://govtrackus.s3.amazonaws.com/legislink/pdf/stat/40/STATUTE-40-Pg553.pdf
[164] https://en.wikipedia.org/wiki/Committee_on_Public_Information
[165] https://mppda.flinders.edu.au/records/365
[166] https://en.wikipedia.org/wiki/Motion_Picture_Production_Code
[167] https://files.eric.ed.gov/fulltext/ED235524.pdf

content. [168] The events and debates leading to the Act, of radio stations owned by members of one party refusing airtime to candidates and topics of the other party and of concerns of commercial censorship affecting the outcome of elections, could be ripped right from today's headlines. [169] As just one example, in 1924, the Progressive Party's presidential candidate was barred from a Republican-owned radio station, while AT&T-owned stations barred broadcasts criticizing government officials and were reluctant to allow socialists to speak. [170] Stations were largely left to their own devices to decide what speech they felt was acceptable to society or aligned with their corporate values – a nearly identical situation to today's social platforms.

In an "attempt to forestall further government regulation" the National Association Of Broadcasters developed a Code of Ethics in 1928 to govern acceptable radio content. In 1929 this Code barred anything "which would commonly be regarded as offensive," anything prohibited by the Post Office as "fraudulent, deceptive, or obscene," and any "advertising statements or claims which he knows or believes to be false, deceptive or grossly exaggerated." Health-related advertising was also afforded special review. [171]

Government Acts Through Private Companies

The rise of these new guidelines to govern the new technologies of motion picture and broadcasting reinforced that new technological developments would be subjected to far greater scrutiny than the newspapers that preceded them. In fact, while cases like Near v. Minnesota granted newspapers wide latitude to publish controversial content, [172] radio was held to a wholly different standard. In the late 1920s Los Angeles evangelist Robert Shuler used his radio station KGEF as a radio-era version of Donald Trump's Twitter account, expressing his unvarnished views and openly attacking individuals and organizations with whom he disagreed, eventually succeeding in ousting several city officials. [173] Similar to Trump, he wielded a vast audience, totaling more than 600,000 listeners at the time. [174] Yet despite his vast audience and stature, his outspoken attacks eventually attracted scrutiny by the Federal Radio Commission that regulated radio licenses and his radio license was revoked, silencing him. [175] A century before Donald Trump's great social silencing, the government's similarly centralized power over the airwaves serves as a reminder of the dangerous power of monopoly control.

At the same time, those who embrace such censorship and advocate the use of radio's historical "public interest" barometer [176] to silence speech would do well to reflect on the ways presidential administrations have leveraged government control over broadcasting to stifle dissent. President Franklin Delano Roosevelt's administration leaned heavily on government control of radio to remind broadcasters they

[168] https://repository.uchastings.edu/cgi/viewcontent.cgi?article=1790&context=hastings_comm_ent_law_journal
[169] https://repository.uchastings.edu/cgi/viewcontent.cgi?article=1790&context=hastings_comm_ent_law_journal#page=9
[170] https://repository.uchastings.edu/cgi/viewcontent.cgi?article=1790&context=hastings_comm_ent_law_journal#page=9
[171] https://books.google.com/books?id=DILNAAAAMAAJ&pg=PA1735
[172] https://en.wikipedia.org/wiki/Near_v._Minnesota
[173] https://digital.library.unt.edu/ark:/67531/metadc798063/m2/1/high_res_d/1002773405-Orbison.pdf
[174] https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history
[175] https://digital.library.unt.edu/ark:/67531/metadc798063/m2/1/high_res_d/1002773405-Orbison.pdf
[176] https://www.brookings.edu/research/revisiting-the-broadcast-public-interest-standard-in-communications-law-and-regulation/

had a "legal duty" to support the White House's policy proposals by virtue of using airwaves "loaned to them temporarily by the government." [177] Thirty years later, President Richard M. Nixon demanded the silencing of "unfair network news coverage" through reinforcing "press objectivity" and a "public re-examination of the role of the media in American life." [178] As media criticism of the administration's Vietnam policies intensified, the White House proposed new regulations that would place stations' support for government policy at the center of their license renewals. [179] Nixon's push to "re-examine" the role of media to prevent what he viewed as harmful coverage bears eerie parallels with Rep. Alexandria Ocasio-Cortez's proposal half a century later to "figure out how we rein in our media environment so that you can't just spew disinformation and misinformation." [180] It seems the more things change the more they stay the same.

Redefining "News" In The Era Of Radio

While it attracted the ire of government, radio's early days also prompted an existential battle with newspapers. Radio's realtime nature and large geographic reach proved formidable competition to the daily cadence of newspapers. Concerned that radio would replace them, newspapers moved to restrict the ability of radio to carry news in a series of battles known as the Press-Radio War, [181] culminating in an uneasy truce with the Biltmore Agreement [182] in which radio largely forfeited the right to report on the news. Radio soon found its way around these limitations and the two mediums eventually coexisted. The end result is that "news" evolved from words on a page into providing information across multiple mediums. [183] As the news media seeks to reinvent itself today in a world in which social media now acts as gatekeeper to the news, there is much the industry could learn by reflecting on this century-old reimagination of just what "news" is.

In 1941 the United States Office of Censorship [184] was created to enforce wartime restrictions on information, ranging from the weather to presidential travel, reinforcing the nation's acceptance of speech controls during times of conflict.

The 1940s, like today's social platform debates, were a period of reflection in the role radio played in society. Concern over the fate of local content, advertiser influence over the news and the state of advertising in general and the role of radio in society were major issues during this period. The FCC's Blue Book, [185] [186] Mayflower Doctrine [187] and Fairness Doctrine [188] [189] all came about during this period in attempts to reshape radio's influence on society.

---

[177] https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history
[178] https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history
[179] https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history
[180] https://www.dailymail.co.uk/news/article-9149219/AOC-slammed-suggestion-federal-commission-rein-press.html
[181] https://books.google.com/books?hl=en&lr=&id=PXHuUO_UJi4C
[182] https://en.wikipedia.org/wiki/The_Biltmore_Agreement
[183] https://books.google.com/books?hl=en&lr=&id=PXHuUO_UJi4C
[184] https://en.wikipedia.org/wiki/Office_of_Censorship
[185] https://en.wikipedia.org/wiki/Blue_Book_(FCC)
[186] https://repository.upenn.edu/cgi/viewcontent.cgi?article=1770&context=asc_papers
[187] https://en.wikipedia.org/wiki/Mayflower_doctrine
[188] https://books.google.com/books?id=cyfQAAAAMAAJ
[189] https://repository.upenn.edu/cgi/viewcontent.cgi?article=1770&context=asc_papers

Television & The Limits Of Political Censorship

The rise of television in the 1950s led to the National Association of Radio and Television Broadcasters Television Code in 1952 in an attempt to avoid "the shadow of incipient censorship by Government regulation." [190] The code mandated, among other things that "law enforcement shall be upheld and the officers of the law are to be portrayed with respect and dignity," that shows airing when children were home emphasize "respect for parents, for honorable behavior and for the constituted authorities of the American community" and that they "foster and promote the commonly accepted moral, social and ethical ideals characteristic of American life." [191] As with radio before it, television was to present a carefully sanitized version of society that excluded perspectives and beliefs viewed as undesirable.

It was also during this period that the courts were called upon ever further to clarify the degree to which broadcasters could or should silence political speech they found offensive in order to protect society. In 1959 the Supreme Court confirmed that broadcasters could not censor a federal candidate's speech even if it was potentially libelous. If they could, "a station so inclined could intentionally inhibit a candidate's legitimate presentation under the guise of lawful censorship of libelous matter." [192] Moreover, "because of the time limitation inherent in a political campaign, erroneous decisions by a station could not be corrected by the courts promptly enough to permit the candidate to bring improperly excluded matter before the public." [193]

Such concerns bear eerie resemblance to the debate over Facebook and Twitter's silencing of the New York Post's Hunter Biden story in October 2020 in the weeks prior to the presidential election. [194] Shortly after its publication, Facebook immediately "reduced its distribution … to reduce the spread of misinformation" [195] until fact checking websites could review it, while Twitter simply banned all links to the story under an ever-changing set of rationales. [196] It took two whole days for Twitter to eventually allow sharing of the story, [197] by which point the combined blackout had had a significant effect on the story's visibility. [198] Today social media platforms and fact checking organizations have normalized the idea of private companies censoring political speech based on what they believe to be true, false or even merely "missing context." The lack of the broadcasting era's government-enforced minimum rights means social platforms are now free to intervene in political speech at will.

Just over a decade later, in 1972 the FCC grappled with the question of when a political candidate's inflammatory speech crosses the line into inciting imminent violence and the power of broadcasters to refuse to air such speech. When self-declared "white racist" Democratic senatorial candidate J.B. Stoner wanted to run radio and television advertisements filled with racial epithets and attacks on African Americans, Atlanta's mayor issued an executive order requesting broadcasters refuse the ads. The NAACP

---

[190] https://books.google.com/books?id=PwzRAAAAMAAJ&pg=PA8
[191] https://books.google.com/books?id=PwzRAAAAMAAJ&pg=PA8
[192] https://supreme.justia.com/cases/federal/us/360/525/
[193] https://supreme.justia.com/cases/federal/us/360/525/
[194] https://www.theverge.com/2020/10/14/21515972/facebook-new-york-post-hunter-biden-story-fact-checking-reduced-distribution-election-misinformation
[195] https://twitter.com/andymstone/status/1316395902479872000
[196] https://www.realclearpolitics.com/articles/2020/10/16/twitter_facebook__hunter_biden_big_tech_as_big_brother_144467.html
[197] https://www.nytimes.com/2020/10/16/technology/twitter-new-york-post.html
[198] https://www.washingtonpost.com/technology/2020/10/15/facebook-twitter-hunter-biden/

and other organizations asked the FCC to issue a waiver to its rules on candidate speech on the grounds that Stoner's remarks were leading to threats of violence. Instead, the FCC upheld its ban on candidate censorship, arguing that this: [199]

> *would amount to an advance approval by the Commission of licensee censorship of a candidate's remarks … Despite your report of threats of bombing and violence, there does not appear to be that clear and present danger of imminent violence which might warrant interfering with speech which does not contain any direct incitement to violence. A contrary conclusion here would permit anyone to prevent a candidate from exercising his rights under [FCC regulations] by threatening a violent reaction.*

Half a century later, this standard of "imminent violence" was cited by social media companies in banning then-President Trump. [200] Twitter justified its removal of the president by citing that "our public interest policy — which has guided our enforcement action in this area for years — ends where we believe the risk of harm is higher and/or more severe." [201] Facebook said that "this is an emergency situation and we are taking appropriate emergency measures, including removing President Trump's video. We removed it because on balance we believe it contributes to rather than diminishes the risk of ongoing violence." [202] Seen through this light, the silencing of Trump by major social media platforms was justified under this historical precedent, since his accounts were not suspended until after the violent storming of the capitol had occurred.

In effect, broadcasters as private companies were left to make the day-to-day decisions of who and what was said on the airwaves, while the government protected selected political candidate speech, seeing it as a bedrock of democracy.

A Changing Society Prompts A Battle Over Who Decides Acceptable Speech

The rapid social change of the 1950s brought renewed attention to the tension inherent in a diverse and changing society over who determines what constitutes acceptable speech. The 1959 Congressional hearing on "Obscene Matter Sent Through The Mail" [203] offers a vivid contemporary glimpse at the debates of the era that could be ripped verbatim from today. Differing perspectives between coastal states and those of the Midwest leading to conflicts over whose standards should prevail. The argument that the lowest denominator of speech must win because the mere existence somewhere of speech one disagrees with is sufficient to cause harm. The conclusion that such harmful speech is the root of all of society's problems and that censorship can solve them. Questions of how to measure the "intent" of a speaker or the "impact" of a given statement. The belief that those wishing to censor public speech "largely reflect the moral climate of the public itself." Even the role of data brokers and personalized advertising. [204] Change "post office" to "social media" and the hearing could have been held in 2021.

---

[199] https://repository.uchastings.edu/cgi/viewcontent.cgi?article=1288&context=hastings_comm_ent_law_journal#page=11

[200] https://www.nbcnews.com/tech/social-media/facebook-youtube-twitter-remove-video-trump-amid-chaos-capitol-n1253157

[201] https://twitter.com/TwitterSafety/status/1346970433049022471

[202] https://twitter.com/guyro/status/1346950532372393985

[203] https://books.google.com/books?id=fYYtAAAAMAAJ&pg=PA232#v=onepage&q&f=false

[204] https://books.google.com/books?id=fYYtAAAAMAAJ&pg=PA232#v=onepage&q&f=false

Two years later, FCC Chairman Newton Minow's 1961 "Vast Wasteland" speech [205] reinforced the similarity between the concerns of the television era and those of today's social platforms, from their impact on children and the loss of local perspectives to the focus on maximizing viewership at all costs and the impact on news. [206]

The Post Office's Battle Against Gay Rights & The Dangers Of Gatekeepers

As American society continued to liberalize in the 1950s, the mainstream press was slow to embrace gay rights. The New York Times ran headlines like "Perverts Called Government Peril," [207] "Hill and Wherry Study Hears There Are 3,500 Deviates in Government Agencies," [208] "Federal Vigilance On Perverts Asked" [209] and "126 Perverts Discharged." [210] Yet while in the eyes of the Times and other mainstream press, gays were "perverts" and "deviants" to be hunted down and excluded from society, there was a vibrant and rapidly growing landscape of gay newspapers, magazines and other publications distributed nationally through the Post Office. [211]

The dependence of these publications on the Post Office monopoly for their distribution made them uniquely vulnerable to censorship under the Comstock Laws and related powers. Publishers could print their materials, but the Post Office decided whether each issue was acceptable speech and could simply refuse to distribute it. The Post Office embraced its role as moral gatekeeper for the nation, [212] leveraging its centralized censorship powers to harass, hinder and even bankrupt gay publications during the 1950s and 1960s. [213]

It was the First Amendment and the court system that eventually overcame the Post Office's censorship of gay publications, [214] but the ability of the Post Office for so many years to silence gay publications offers a poignant reminder of the dangers of the kinds of central gatekeepers that social platforms now play in the digital world.

The "Intractable" Problem Of "Acceptable Speech"

The growing visibility and voice of traditionally underrepresented communities required increasing intervention by the US Supreme Court to help define just what precisely constituted harmful speech.

---

[205] https://www.americanrhetoric.com/speeches/newtonminow.htm
[206] https://www.forbes.com/sites/kalevleetaru/2019/01/08/minows-wasteland-how-the-webs-problems-are-those-of-television-half-a-century-ago/
[207] https://www.nytimes.com/1950/04/19/archives/perverts-called-government-peril-gabrielson-gop-chief-says-they-are.html
[208] https://www.nytimes.com/1950/05/20/archives/inquiry-by-senate-on-perverts-asked-hill-and-wherry-study-hears.html
[209] https://www.nytimes.com/1950/12/16/archives/federal-vigilance-on-perverts-asked-senate-group-says-they-must-be.html
[210] https://www.nytimes.com/1952/03/26/archives/126-perverts-discharged-state-department-reports-total-ousted-since.html
[211] https://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1530&context=wmjowl
[212] https://www.vox.com/2014/5/28/5756494/the-homophobic-history-of-the-post-office
[213] https://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1530&context=wmjowl
[214] https://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1530&context=wmjowl

In its 1957 Roth v. United States ruling, [215] the court determined that "the standard for judging obscenity, adequate to withstand the charge of constitutional infirmity, is whether, to the average person, applying contemporary community standards, the dominant theme of the material, taken as a whole, appeals to prurient interest." It reiterated that "the protection given speech and press was fashioned to assure unfettered interchange of ideas for the bringing about of political and social changes desired by the people" and that "all ideas having even the slightest redeeming social importance - unorthodox ideas, controversial ideas, even ideas hateful to the prevailing climate of opinion - have the full protection of the guaranties."

It also pushed back on censors' race to the bottom in which any speech that offended anyone could be banned, ruling that the State of Michigan could not "reduce the adult population of Michigan to reading only what is fit for children." [216]

The civil rights era brought with it a growing effort by southern states to silence the national debate around civil rights. In an effort to "caus[e] reckless publishers of the North … to make a re-survey of their habit of permitting anything detrimental to the South and its people to appear in their columns," southern states had filed more than $300 million in libel suits against news outlets by 1964. Their efforts had a chilling effect. By the time of the Supreme Court's New York Times Co. v. Sullivan ruling that year, [217] the New York Times had withdrawn all of its reporters from Alabama for a year and CBS was prepared to cease all coverage of the southern civil rights movement. [218]

The ability of southern states to use libel law to silence the public debate reminds us that in the battles over free speech, even the most straightforward seeming rules will be repurposed to silence dissenting voices. Notably, this same tactic was resurrected a few decades later in the early days of the web as a way for large companies to silence online criticism. [219]

Over the coming decade the Supreme Court continued to wrestle with just how to define what speech was allowable and not allowable, offering a stark reminder that the goals of social platforms in precisely defining their "community guidelines" have long remained elusive. It was during this era that the Court reminded us that the "right to receive information and ideas, regardless of their social worth is fundamental to our free society." [220] That in contrast to today's increasing push towards "majority rule" speech guidelines, the Constitution's "guarantee is not confined to the expression of ideas that are conventional or shared by a majority … And, in the realm of ideas, it protects expression which is eloquent no less than that which is unconvincing."

In an era in which Amazon bans books and streaming services remove shows in order to protect society from what they see as harmful content consumed in one's home, it is worth reflecting on the Court's 1969 counsel on this question: [221]

[215] https://supreme.justia.com/cases/federal/us/354/476/
[216] https://supreme.justia.com/cases/federal/us/352/380/
[217] https://en.wikipedia.org/wiki/New_York_Times_Co._v._Sullivan
[218] https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=11530&context=journal_articles#page=5
[219] https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1071&context=dlj
[220] https://supreme.justia.com/cases/federal/us/394/557/
[221] https://supreme.justia.com/cases/federal/us/394/557/

> *Whatever may be the justifications for other statutes regulating obscenity, we do not think they reach into the privacy of one's own home. If the First Amendment means anything, it means that a State has no business telling a man, sitting alone in his own house, what books he may read or what films he may watch. Our whole constitutional heritage rebels at the thought of giving government the power to control men's minds. And yet, in the face of these traditional notions of individual liberty, Georgia asserts the right to protect the individual's mind from the effects of obscenity. We are not certain that this argument amounts to anything more than the assertion that the State has the right to control the moral content of a person's thoughts. To some, this may be a noble purpose, but it is wholly inconsistent with the philosophy of the First Amendment.*

At the time the Court wrote these words, few could imagine a world half a century later in which one company now accounts for as much as 72% of adult new book sales in the country [222] or a handful of platforms acting as gatekeepers controlling access to much of the entertainment and informational world. The powers of absolute censorship once afforded only to the State are now in the hands of private companies that have normalized all of the things the Court warned against.

As social platform "community standards" [223] increasingly trend towards a lowest common denominator that bar speech that might offend someone somewhere, it is worth reflecting on the Court's view half a century ago in Miller v. California: [224]

> *The idea that the First Amendment permits government to ban publications that are 'offensive' to some people puts an ominous gloss on freedom of the press. That test would make it possible to ban any paper or any journal or magazine in some benighted place. The First Amendment was designed 'to invite dispute,' to induce 'a condition of unrest,' to 'create dissatisfaction with conditions as they are,' and even to stir 'people' to anger.' The idea that the First Amendment permits punishment for ideas that are 'offensive' to the particular judge or jury sitting in judgment is astounding. No greater leveler of speech or literature has ever been designed. To give the power to the censor, as we do today, is to make a sharp and radical break with the traditions of a free society. The First Amendment was not fashioned as a vehicle for dispensing tranquilizers to the people. Its prime function was to keep debate open to 'offensive' as well as to 'staid' people. … the materials before us may be garbage. But so is much of what is said in political campaigns, in the daily press, on TV, or over the radio. By reason of the First Amendment—and solely because of it—speakers and publishers have not been threatened or subdued because their thoughts and ideas may be 'offensive' to some.*

Justice John Marshall Harlan noted at the time that deciding just what constitutes acceptable speech "is almost intractable" [225] [226] and "has produced a variety of views among the members of the Court unmatched in any other course of constitutional adjudication." [227] Indeed, the long, complicated history

[222] https://www.wsj.com/articles/they-own-the-system-amazon-rewrites-book-industry-by-turning-into-a-publisher-11547655267
[223] https://www.facebook.com/communitystandards/
[224] https://supreme.justia.com/cases/federal/us/413/15/
[225] https://books.google.com/books?id=9LHQehQel4UC&pg=PR9
[226] https://books.google.com/books?id=aQI0Ee1JfqQC&pg=PT86
[227] https://books.google.com/books?id=tanODwAAQBAJ&pg=PA197

of First Amendment litigation and legislation reminds us just how immensely complex and fluid these issues are, with few easy solutions. [228]

The Court's ultimate recommendation for how to decide the rules of acceptable speech for society? To put it to a vote of the people, rather than leave it with the courts: [229]

> *If there are to be restraints on what is obscene, then a constitutional amendment should be the way of achieving the end. There are societies where religion and mathematics are the only free segments. It would be a dark day for America if that were our destiny. But the people can make it such if they choose to write obscenity into the Constitution and define it. We deal with highly emotional, not rational, questions. To many the Song of Solomon is obscene. I do not think we, the judges, were ever given the constitutional power to make definitions of obscenity. If it is to be defined, let the people debate and decide by a constitutional amendment what they want to ban as obscene and what standards they want the legislatures and the courts to apply. Perhaps the people will decide that the path towards a mature, integrated society requires that all ideas competing for acceptance must have no censor. Perhaps they will decide otherwise. Whatever the choice, the courts will have some guidelines. Now we have none except our own predilections.*

It seems these lessons of the past have long ago been forgotten. In the place of such a democratic vision, Facebook today has appointed its own "Supreme Court" [230] comprised of 20 judges unaccountable to the American public. [231] In fact, three quarters of the judges with absolute authority over the national debate that plays out on Facebook each day are now from outside the US. [232] Former foreign elected officials and the citizens of other nations now have oversight over the speech of the president of the United States. [233] From putting speech to a vote of the American public to outsourcing it to former officials and citizens from other countries, the web has globalized what was previously a national debate.

Cable Television & The Gatekeepers

Fast forward a decade from these cases and the rise of cable television and its accompanying ability to host hundreds of channels rendered moot the justification that government control of the airwaves was a necessary correlate of their scarcity. Suddenly there was room for every viewpoint. Yet, as Michael Pollan outlined in 1981: [234]

> *…although cable does promise great numbers of [viewpoints], the gardens in which they flourish will be owned by a few media giants. Each cable system is a municipally licensed monopoly; even when there are 104 different channels, all of these are ultimately controlled by a single company. Should a cable company be free to dominate the political contents of its 104 channels when we do not want a network to dominate even one? … Rather than decentralize media power, the television revolution may end up concentrating far more of it in the same few hands - hardly an environment for political diversity.*

---

[228] https://www.law.cornell.edu/constitution-conan/amendment-1/freedom-of-expression-speech-and-press
[229] https://www.law.cornell.edu/supremecourt/text/413/15
[230] https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court
[231] https://oversightboard.com/news/327923075055291-announcing-the-first-members-of-the-oversight-board/
[232] https://oversightboard.com/news/327923075055291-announcing-the-first-members-of-the-oversight-board/
[233] https://www.bbc.com/news/technology-56781104
[234] https://www.nytimes.com/1981/12/22/opinion/keeping-television-regulated.html

To curtail this consolidation, Pollan proposed an antidote of "vigorous antitrust enforcement, strengthened prohibitions on cross ownership by conglomerates, and the designation of cable as a 'common carrier' required to lease channels to anyone who could pay." [235]

The Rise Of Monopoly Gatekeepers

At the same time, the argument could be made that by virtue of the government monopolies over the Post Office and airwaves, both publishing and broadcasting had long been gatekept mediums, in which a vibrant landscape of divergent voices across the nation were forced to conform to a single set of rules devised by the federal government in order to be distributed. Thus, the rise of cable monopolies as gatekeepers represented not a new phenomenon, but rather the continued growth of gatekeeping over the nation's 200-year history.

The difference is that the rise of broadcasting came alongside the strengthening of protections for newspapers. While radio and television fell under strict government supervision and control over their speech, newspapers were largely exempt from these rules. [236] This meant that what the government prohibited broadcasters from saying, newspapers could publish, ensuring that while entire swaths of society were still excluded, no single entity wielded power over the entirety of the national discourse.

Over time, the consolidation of newspapers has steadily eroded this viewpoint diversity. In 1920, 42.6% of cities with daily newspapers had two or more competing daily papers. By 1940, this had fallen to 12.7%, collapsing to 4.2% in 1960 and just 1.9% by 1986. This rise of monopoly papers coincided with the loss of newspaper independence as papers were folded into national chains. In 1920, 92% of newspapers in the US were independent, falling to 68.2% in 1960, 30.1% in 1986 and just 24.8% in 1996. [237] In the 1930s as newspapers sought to own radio stations, Joy Elmer Morgan, editor of the Journal of the National Educational Association articulated the growing concern that "if monopoly is bad in the material realm it is infinitely worse in the realm of instruments for the formation of public opinion." [238] By 1960, New Yorker press critic A. J. Liebling argued that "diversity – and the competition that it causes – does not insure [sic] good news coverage, but it increases the chances." [239] Yet by the 1990s the focus of newspapers had shifted from reporting to "the responsibility to produce a return for our shareholders." [240]

With the consolidation of newspapers, their ability to counter government control of broadcasting steadily weakened as control over broadcasting speech continued to centralize, with cable monopolies arriving just as the press had reached a critical point in consolidation.

For their part, cable companies still wield enormous power over television distribution today. When Democrats wanted to silence Fox News, [241] [242] they had no formal governmental power to ban it. Instead,

[235] https://www.nytimes.com/1981/12/22/opinion/keeping-television-regulated.html
[236] https://en.wikipedia.org/wiki/Near_v._Minnesota
[237] https://transition.fcc.gov/osp/inc-report/INoC-1-Newspapers.pdf#page=4
[238] https://books.google.com/books?hl=en&lr=&id=PXHuUO_UJi4C
[239] https://books.google.com/books?hl=en&lr=&id=PXHuUO_UJi4C
[240] https://transition.fcc.gov/osp/inc-report/INoC-1-Newspapers.pdf#page=6
[241] https://www.wsj.com/articles/just-asking-for-censorship-11614295623
[242] https://www.wsj.com/articles/the-censorship-party-11614296803

joining with outlets like CNN, [243] they publicly called on cable providers to voluntarily remove it, since doing so would sharply limit its accessibility.

Four decades later, this consolidation of power has played out to its logical conclusion, with just a handful of social media companies now wielding enormous power over the digital world, acting as the ultimate gatekeeper to everything from entertainment to news. [244] It also reminds us that the ultimate power over information lies not with content creators, but with the gatekeepers that control the flow of that information. As Pollan noted, expenses aside, anyone could create a television station. [245] The problem was that access to the citizenry was controlled by monopolies that could simply decide not to carry that station. So too today can anyone create a website somewhere on the internet, while the access to the public through search engines and social platforms are controlled by a handful of companies that are under no obligation to include it.

For example, Sci-Hub [246] is a website that provides illegal access to millions of academic papers in violation of copyright law. [247] Despite the illegality of its content, the site has continued to grow over nearly a decade simply by transitioning to new domain names and web hosting as old ones are seized or shut down. While some internet providers have blocked access to the site for their users, the decentralized nature of the web at large means there is little publishers can do to restrict access to it globally. In contrast, Twitter simply deleted the site's Twitter account worldwide at the request of a single government. [248]

This gatekeeping extends even to the plumbing of the modern digital world. When Twitter competitor Parler attracted substantial attention, the two dominate mobile phone operating systems banned it and cloud computing vendors barred it from using their services, effectively wiping it from the digital world almost overnight. [249]

The Web And Section 230

Like motion pictures, radio and television before it, the web made it possible to reach ever more people at ever greater speed. Unlike the technologies that preceded it, its avoidance of government-controlled infrastructure like airwaves or the Post Office and its digital nature meant existing state and federal censorship regimes did not readily apply. Yet, like the introduction of all the technologies that preceded it, government was eager to step forward and define the concept of "obscenity" and acceptable speech for this new medium.

The Communications Decency Act of 1996 [250] brought the regulation of the broadcast era to the web, criminalizing the transmission to minors of "obscene or indecent" material and content "patently offensive as measured by contemporary community standards, sexual or excretory activities or organs."

[243] https://www.cnn.com/2021/01/08/media/tv-providers-disinfo-reliable-sources/index.html
[244] https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/
[245] https://www.nytimes.com/1981/12/22/opinion/keeping-television-regulated.html
[246] https://sci-hub.se/
[247] https://en.wikipedia.org/wiki/Sci-Hub
[248] https://sci-hub.se/
[249] https://www.realclearpolitics.com/articles/2021/01/12/the_great_social_silencing_145014.html
[250] https://en.wikipedia.org/wiki/Communications_Decency_Act

This was later nullified by the Supreme Court, [251] but taking a chapter from the Comstock era's focus on regulating content through access points, the Children's Internet Protection Act later reestablished some of these protections. [252]

Today the Communications Decency Act is most famous for Section 230, [253] also known as "the 26 words that created the Internet." [254] Its most important provision states that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." [255] In short, by exempting internet companies like social media platforms from legal liability for libel or other harms conducted on their platforms, the web was freed from the speech rules that constrained broadcasters.

Yet Section 230 went further to not just exempt companies from harmful speech on their platforms, but to exempt them from claims related to the removal of content, stating "no provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." It is this clause that strips from the web the traditional concepts of free speech as enshrined by the First Amendment. Companies can remove whatever they themselves view as "objectionable" without recourse or accountability.

In today's debates over social media's role in society, Democrats most commonly object to the first clause, arguing that platforms should have greater legal responsibility for removing what Democrats view as harmful speech. In contrast, Republicans most commonly cite the second clause as the most problematic, arguing that its protections enable platforms to silence legitimate political discourse with impunity.

Concluding the nation's long march away from state and local influence over what constitutes acceptable speech, Section 230 forbids cities and states from enforcing any additional speech restrictions: "No cause of action may be brought and no liability may be imposed under any State or local law that is inconsistent with this section."

Section 230 has not remained immutable. In 2018, it was amended via the Stop Enabling Sex Traffickers Act and Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA-SESTA) bill [256] to add language that explicitly revoked immunity regarding violations of federal and state sex trafficking laws. The passage of FOSTA-SESTA provides a clear legal precedent for selectively modifying Section 230 over time to exempt other forms of speech from protection.

[251] https://supreme.justia.com/cases/federal/us/521/844/
[252] https://en.wikipedia.org/wiki/Children%27s_Internet_Protection_Act
[253] https://www.law.cornell.edu/uscode/text/47/230
[254] https://www.amazon.com/Twenty-Six-Words-That-Created-Internet/dp/1501714414/ref=sr_1_1
[255] https://www.law.cornell.edu/uscode/text/47/230
[256] https://www.congress.gov/bill/115th-congress/house-bill/1865/text

CHALLENGES & SOLUTIONS

How can we prevent the dystopian world that opens this report from becoming reality?

Reflecting on America's two-century experiment with speech censorship, the lessons learned from the early days of the web and the state of social media today, what are some of the greatest challenges social platforms and their moderation practices pose to society today? Which of these challenges are new and which are merely the same quandaries that have emerged with the introduction of each new communications technology since America's founding?

With each new communications medium comes a change in the rules and norms of societal behavior. As Gwenyth Jackaway describes this process: [257]

> *every culture has written and unwritten rules governing the flow of information in society. These are the rules of social discourse: rules that cover who should speak to whom about what; rules about what should be said, the way it should be said and the circumstances in which it should be said; and rules about who should have and control access to information and which sources of information are considered legitimate. … New media can disrupt these established patterns of communication. With their capacity to transmit and receive information in new ways, new media often render the old rules obsolete or impossible to enforce. … With their ability to send information through new channels, in new ways, at greater speeds with higher efficiency, new media demand that we alter our familiar ways of communicating with each other.*

As this same process has played out with the evolution of social media, what are a few of the representative challenges the current social landscape poses to society and what are some possible legislative, educational and technical solutions?

Accountability & Transparency

For all the concern over "community guidelines," content moderation, fact-checking and advertising policies, we have few of the actual data points necessary to evaluate how well current approaches to content moderation and combatting falsehoods are working. Could it be that the public would actually agree with their decisions most of the time and it is just a few high-profile mistakes that are feeding the public debate over their censorship powers? Conversely, are the companies getting it wrong much more than we, or even they, realize? [258]

On paper, the platforms' content moderation practices and fact-checking partnerships seem like reasonable solutions to the difficult task of keeping bad actors from disrupting their digital communities.

[257] https://books.google.com/books?hl=en&lr=&id=PkXTrh4eDlcC

[258] https://www.forbes.com/sites/kalevleetaru/2016/05/13/is-facebooks-trending-topics-biased-against-africa-and-the-middle-east/

Yet how closely do the companies adhere to these rules in practice? To what degree do the unconscious biases of the companies' engineers manifest themselves in their algorithms? [259]

The companies themselves openly acknowledge the difficulty of their work. Facebook's Head Of Product Policy Monika Bickert noted in 2017 that its "policies do not always lead to perfect outcomes. That is the reality of having policies that apply to a global community where people around the world are going to have very different ideas about what is OK to share. I'll be the first to say that we're not perfect every time." [260]

Would the American public be as supportive of Facebook's decisions and of content moderation more generally if they understood the disproportionate ways it can impact underrepresented voices or the unevenness in how the platforms apply their rules? [261] Would the public have supported Facebook's policy of allowing graphic threats of violence against women, [262] gender-based attacks on women drivers, [263] race-based attacks on black children, [264] providing a special marketing category for "Jew haters" [265] or allowing its recommendation algorithms to encourage anti-Semitism? [266]

None of these insights were provided by the companies themselves – they were all leaked or discovered by researchers outside the company shining light on its practices. Yet their centrality in modern life means we cannot depend on these chance revelations; companies must be compelled to provide sufficient transparency to enable public debate over their policies.

In order to accurate examine the impact of social platforms on society, we need data that captures the daily functioning of our modern public squares.

Legislative Solutions

Social media platforms today have no legal obligation to provide even the most basic of transparency around their policies and how they enforce them, how they train their algorithms and their known biases or any of the critical details that would help the public evaluate their impact on society. Section 230 could be amended to require that in exchange for its safe harbor protections, companies are required to provide both clarity and transparency around their content moderation decisions and the algorithms that power them, along with their growing use of sensitive user data for research.

Auditing Datasets

---

259 https://www.realclearpolitics.com/articles/2020/07/12/facebook_audit_exposes_algorithm_biases_in_policing_speech.html

260 https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

261 https://www.cnn.com/2019/11/01/health/facebook-harassment-eprise/index.html

262 https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

263 https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

264 https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

265 https://www.theguardian.com/technology/2017/sep/14/facebook-advertising-jew-hater-antisemitism

266 https://themarkup.org/news/2020/11/24/facebook-ban-holocaust-deniers-antisemitism

As outlined below under Technical Solutions, social platforms should be required to publish a collection of datasets that would permit external auditing of their moderation activities, enabling scholarly and societal scrutiny of the decisions that increasingly govern the digital public square.

Clarity Regarding Research Data Practices

In addition to its public-facing platform, most social media platforms have active research programs that conduct experiments on their users in the name of science. From secretly manipulating the emotions of its users [267] to predicting when children as young as 14 are feeling clinically depressed in order to maximize their malleability to advertising, [268] to harvesting medical records, [269] to permitting mass extraction of user data for campaigning, [270] to its latest initiative to make its users' aggregated data available for academic research to the world, [271] Facebook in particular has been a leader in this area.

New legislation is needed to clarify the rights of users to not be experimented upon, especially children. Existing laws like GDPR largely exempt such research from most regulation, [272] meaning dedicated laws are needed that explicitly target experimentation and data access by social platforms for any purpose other than operating their platforms. At the very least, users should be granted the basic right to "opt-out" from research, [273] which has long been enshrined in the US Department of Health and Human Services Policy for the Protection of Human Research Subjects - the so-called "Common Rule." Even the top scientific journals now largely view the Common Rule as not applying to social media, [274] underscoring the urgency of new protections.

Detailed Plain-English Explanation Of All Enforcement Actions

Social media companies routinely delete posts and suspend or ban users without any explanation or by citing vague or unrelated policies. Search the web for the phrase "suspended with no explanation" along with the name of any major social platform and endless pages of forums detailing user experiences will be returned. Even for high-profile enforcement actions, the explanation can even change over time. Twitter originally claimed it was banning sharing of the New York Post's Hunter Biden story because it was "harmful." It then said it was a violation of its hacked materials policy, before changing its story a third

[267] http://www.pnas.org/content/111/24/8788.full
[268] http://www.theaustralian.com.au/business/media/digital/facebook-targets-insecure-young-people-to-sell-ads/news-story/a89949ad016eee7d7a61c3c30c909fa6
[269] https://www.forbes.com/sites/kalevleetaru/2018/04/05/facebooks-medical-research-project-shows-it-just-doesnt-understand-consent/
[270] https://www.forbes.com/sites/kalevleetaru/2018/03/19/why-are-we-only-now-talking-about-facebook-and-elections/
[271] https://www.forbes.com/sites/kalevleetaru/2018/04/12/is-facebooks-new-academic-initiative-even-more-frightening-than-its-own-research/
[272] https://www.forbes.com/sites/kalevleetaru/2019/05/23/gdprs-massive-research-exemption-facebook-and-social-science-one/
[273] https://www.forbes.com/sites/kalevleetaru/2019/02/02/what-does-it-mean-to-consent-to-the-use-of-our-data-in-the-facebook-era/
[274] https://www.forbes.com/sites/kalevleetaru/2018/08/13/social-science-one-and-how-top-journals-view-the-ethics-of-facebook-data-research/

time to say it violated its personal information policy. [275] After a public outcry it finally removed the ban and admitted it was "wrong" [276] and "a total mistake." [277]

Without limiting their ability to perform censorship under Section 230, internet companies should be required to clearly document in plain English the rationale behind each enforcement action. Such explanations should cite the specific policy and provide at least a paragraph or more of text clearly explaining why the moderator believed the post to be a violation. Requiring moderators to clearly explain their decision rather than simply clicking "keep" or "remove" forces them to carefully reason about why the believe the post violates the policy. Most importantly, it creates a documentation trail that users can optionally share with external researchers to evaluate the consistency of the company's decisions and how accurately it is implementing its policies.

This documentation requirement must extend to the algorithmic content moderation companies are increasingly relying upon. Today's algorithms are black boxes that the companies themselves don't fully understand and offer myriad opportunities for inadvertent bias and error. For example, when Twitter accidently banned all mention of the city of Memphis in March 2021, [278] the company would likely have caught the error far sooner if it had been required to explain to each user why it believed their tweet mentioning the city was a violation of its policies.

Companies would likely argue that requiring such explanations for all content would be cumbersome and costly, but trust in the company's moderation policies is essential to trust in our online public squares.

In addition, companies should be required to offer any user whose account or posts are subject to enforcement action and who disagrees with the outcome with the option of a live chat with a real human moderator to appeal the decision, with a guaranteed turnaround time of less than 12 hours and a bias towards restoring the post.

Clear Rules With Precise Definitions And Equal Enforcement

Many of the disagreements with social moderation decisions come from a lack of clear rules defining precisely what is permitted, preventing open societal debate about the acceptability of those rules, along with the uneven enforcement of those rules. Precisely defining the rules of today's internet platforms would permit a more informed societal debate and make it easier for users to understand whether their speech complies with those rules. For example, Microsoft prohibits the use of its Office 365 software to "engage in activity that is harmful to you … or others … [including] communicating hate speech." [279] Yet asked what it considered to be "harmful" or "hate" speech and how many users it had banned under

[275] https://www.realclearpolitics.com/articles/2020/10/16/twitter_facebook__hunter_biden_big_tech_as_big_brother_144467.html

[276] https://www.cnbc.com/2020/10/16/twitter-ceo-jack-dorsey-says-blocking-post-story-was-wrong.html

[277] https://nypost.com/2021/03/25/dorsey-says-blocking-posts-hunter-biden-story-was-total-mistake/

[278] https://www.theguardian.com/technology/2021/mar/15/twitter-accidentally-blocks-users-who-post-the-word-memphis

[279] https://web.archive.org/web/20160728121715/https:/www.microsoft.com/en-us/servicesagreement/upcoming.aspx

these rules, the company declined to comment. [280] Similarly, Airbnb bars members of "hate groups" but leaves out any detail regarding how it defines such groups. [281]

At the same time, even when the rules are clear, such as barring any protest announcements that don't require social distancing or barring calls for violence, those rules are aggressively enforced for some communities, but quietly waived for others due to political considerations. [282] [283]

Companies are under no legal obligation to clarify their speech polices or even modify their written policies to list exemptions. The first that most of the public knew of Uber and Lyft's social media policies was when the companies banned Laura Loomer over her anti-Muslim tweets. [284]

When asked why they don't provide greater clarity around their speech policies, social platforms have often argued that doing so would help bad actors find loopholes and exceptions. [285] The same is true with America's legal system, in which defendants and their lawyers search for technicalities or exceptions, but we accept that as a cost of an open and transparent legal system.

Congress could modify Section 230 to require that any content moderation that platforms perform must be in accordance with clearly established written rules that outline policies in precise plain English and which are enforced evenly for all users.

Educational Solutions

Increasing accountability and transparency for social platforms through new laws and new datasets requires a society that understands the importance of such transparency. Rather than blindly trusting that social platforms are acting in the best interests of society, it is imperative that the press, public and policymakers learn how to think more critically about platforms' holistic impact on democracy.

Technical Solutions

Creating transparency around social platforms begins with the data necessary to evaluate their actions. Below are a set of ten datasets that Congress could demand from social media companies that would begin to provide the critical insights needed to understand their roles in our modern democracy and highlight areas that may require further legislative action.

Algorithmic Trending Datasets

[280] https://www.forbes.com/sites/kalevleetaru/2019/07/26/censorship-comes-for-the-desktop-how-microsoft-has-infused-values-into-windows-and-office/
[281] https://news.airbnb.com/airbnb-announces-capitol-safety-plan-for-the-inauguration/
[282] https://www.realclearpolitics.com/articles/2020/04/22/facebooks_covid-protest_ban_renews_censorship_concerns_143003.html
[283] https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346
[284] https://www.nbcnews.com/news/us-news/laura-loomer-banned-uber-lyft-after-anti-muslim-tweetstorm-n816911
[285] https://www.forbes.com/sites/kalevleetaru/2019/04/23/twitter-follows-facebooks-dystopian-path-towards-unaccountable-automated-content-filtering/

The power of algorithms to shape our awareness of events around us was driven home in 2014 when Twitter chronicled the unrest in Ferguson, Missouri, [286] while Facebook was filled with the smiling faces of people dumping buckets of ice water over their heads. [287] A public dataset capturing how public posts are being prioritized or deemphasized by these algorithms across classes of users and over time would provide insights into inadvertent biases in these algorithms and provide greater visibility into what the public is and is not seeing.

Automatic Database of Violating Public Posts

Given that all tweets are publicly viewable and already accessible to researchers using Twitter's data APIs (application programming interfaces), there would be few privacy implications in requiring Twitter to provide a public database of all tweets the platform flags each day, along with a description of why Twitter believed each tweet was a violation of its rules or disputed by a fact-checker. Such a database would permit at-scale analysis of the kinds of content Twitter's moderation efforts focus on, while the ability to compare those violating tweets against the rest of Twitter would make it possible to assess how evenhanded the platform's removal efforts are.

The Lumen DMCA takedown database could serve as a model, in which companies publish DMCA and other legal takedown requests (such as court orders to remove content illegal under federal law) to a public searchable website where researchers, policymakers, press and the public can search and examine them. Details like precise URLs of infringing content are restricted from public access (to avoid acting as a search index to illegal content) but all other information is publicly accessible and all details are available to researchers, journalists and others.

For publicly accessible content like tweets, all removed content could be indexed into such a database.

Social companies would likely argue that this would empower bad actors since they could simply point people to the archived copy of the deleted post and it would essentially become the largest misinformation publisher in existence. Yet, here Lumen's preexisting solution to redact full URLs shows that minor tweaks, such as limiting certain details only to journalists and researchers, would avoid this.

Politiwoops [288] already archives in a searchable database tweets by public officials [289] that the individuals themselves later deleted (rather than social media companies removing). While Twitter suspended the project's access in 2015 [290] it eventually restored its access [291] and has allowed it to continue since. However, the archive contains only those tweets that politicians delete themselves, not content that Twitter removes as a violation, though in most cases it will catch such content since Twitter typically does not delete tweets, but rather locks an account until the user deletes it themselves. However, even in those cases, there is no explanation in those cases that it was a forced deletion or details about why Twitter felt it was a violation.

[286] https://www.usatoday.com/story/tech/2014/09/02/facebook-twitter-ferguson-icebucketchallenge/14818505/
[287] https://www.theguardian.com/commentisfree/2014/aug/19/ice-bucket-challenge-ferguson-social-media
[288] https://projects.propublica.org/politwoops/
[289] https://projects.propublica.org/politwoops/users
[290] https://arstechnica.com/tech-policy/2015/06/political-deleted-tweet-archive-shuttered-by-twitter-over-privacy-expectation/
[291] https://www.reuters.com/article/us-usa-election-politwoops/twitter-to-revive-politwoops-archive-of-politicians-deleted-tweets-idUSKBN0UE16520151231

One could imagine a system in which removed posts by public figures are archived in their entirety ala a Politiwoops model, while removed posts voluntarily submitted by ordinary citizens are archived akin to Lumen, in which the public can see basic details, while journalists and researchers can see all details. Each entry would include the full explanation provided to the user of why the post was removed.

Entries would also include basic demographic details about the poster as self-reported by the user or purchased or inferred by the platform if the user allows. This would include all demographic-related advertising selectors. For example, if the platform allows advertisers to target LGBTQ minorities and those selectors are attached to this user, the user could be asked if they are willing to share those selectors as part of the public record. Some users might not, while others might be glad to share the selectors.

Database of Deleted & Exempted Protest Posts

Protest marches are increasingly being organized over social media. As platforms extend their censorship to these posts, [292] [293] they are able to control speech that occurs beyond their digital borders. This makes understanding how platforms moderate protest-related speech uniquely important. For weeks Facebook touted its removal of COVID "reopening" protests that did not require social distancing, yet quietly waived those rules for the George Floyd protests. [294] Having a centralized database of protest posts removed by platforms as well as those exempted from its rules would go a long way towards understanding how much the platforms are shaping the offline discourse.

Database of Exempted Posts

A common criticism of content moderation is the unevenness with which it is applied. Why do some users seemingly face constant enforcement action while others posting the exact same material face no consequences? Why is one politician's post preserved as "newsworthy" while another is removed as a violation? A critical missing component in our understanding of content moderation is the degree to which companies create silent exemptions from their rules. On paper, Facebook prohibits all forms of sexism, racism, bullying and threats of violence, but in practice, the company allows some posts as "humor" [295] or otherwise declines to take action. [296] How often do users report posts that the company determines are not a violation? And does it systematically exempt certain kinds of content? Compiling a central database of posts the companies rule are not violations would offer critical insights into how evenhanded they are and where their enforcement gaps are.

---

[292] https://www.realclearpolitics.com/articles/2020/01/29/blocked_march_for_life_tweets_raise_free_speech_concerns_142260.html

[293] https://www.realclearpolitics.com/articles/2020/04/22/facebooks_covid-protest_ban_renews_censorship_concerns_143003.html

[294] https://www.realclearpolitics.com/articles/2020/04/22/facebooks_covid-protest_ban_renews_censorship_concerns_143003.html

[295] https://www.realclearpolitics.com/articles/2020/07/02/facebook_boycott_reveals_triad_shaping_the_public_square_143603.html

[296] https://www.cnn.com/2019/11/01/health/facebook-harassment-eprise/index.html

In addition to a database of actual removals, companies should be required to provide for researchers and journalists (potentially with certain redactions) a list of posts that were reported to the platform as a violation and which the platform ultimately determined were not a violation and allowed to remain. [297]

This goes to the heart of one of the most common criticisms of social platforms: double standards. That the exact same post by one user is removed as impermissible speech, but deemed completely fine when written by another.

Database of Fact-Checked Posts

What are the kinds of posts that social platforms delete or flag as having been disputed by fact-checking organizations? Are climate change posts flagged more often than immigration posts? How are platforms managing the constantly changing guidelines for COVID-19, when earlier in the pandemic posts recommending masks would have in theory been a violation of the platforms' "misinformation" rules governing health information that goes against CDC guidance? How often are posts flagged based on questionable ratings [298] or potentially conflicted sources? [299]

In an ideal world, platforms would be required to compile a database of every post they flag as being disputed by a fact-checker. For public posts such as those on Twitter, this would be trivial, but for platforms like Facebook, this would pose a privacy challenge. One possibility would be to require platforms such as Facebook to provide a daily report listing the URL of every fact check they relied upon to flag a user post that day, along with how many posts were flagged based on that fact check. For example, of all of the climate change fact checks published over the years, which are the ones that yield the most takedowns on social platforms? Do the most heavily cited fact checks rely on the same sources of "truth" as other fact checks on that topic or is a particular source, such as an academic "expert," having an outsized influence on "truth" on social platforms? Such data would also help fact-checkers to periodically review their most-cited fact checks to verify that their findings still hold, while during pandemic public health officials could use it to flag emerging contested narratives.

Database of Journalist & Politician Private Post Violations

Most social platforms, such as Facebook and Instagram, are a mixture of public and private content. Publicly shared content violations could be compiled and disseminated to researchers, as could public tweets, but private content such as non-public Facebook posts that are deleted or flagged as misinformation pose unique privacy challenges. One possibility would be to treat the verified official accounts of journalists and elected officials as different from other users, given their outsized role in the public discourse, and to automatically make available to researchers any posts by those accounts that are later deleted as violations of platform rules or disputed by fact-checkers.

A separate voluntary submission database could allow ordinary users to submit their own posts that were deemed violations, along with the explanation they received regarding the violation. Having a single

---

297 https://www.realclearpolitics.com/articles/2020/07/29/heres_the_data_congress_needs_to_regulate_social_media_143824.html

298 https://www.realclearpolitics.com/articles/2020/05/06/mired_in_semantics_fact_checkers_miss_the_covid-19_moment_143125.html

299 https://www.realclearpolitics.com/articles/2020/05/12/facebooks_evidence-free_false_rating_143182.html

centralized database of such removals would make it easier to understand trends in the kinds of content platforms are most heavily policing and whether there is public agreement with the platforms' decisions.

Demographic Database of Content Removals

Social platforms use algorithms to estimate myriad demographic characteristics of their users, including race, gender, religion, sexual orientation and other attributes that marketers can use to precisely target their ads. [300] While these attributes are imperfect, the fact that the companies make them available for ad targeting suggests they believe they are sufficiently accurate to build an advertising strategy upon. The companies should be required to compile regular demographic percentage breakdowns of deleted and flagged posts for each of their community guidelines and fact checks. For example, what percentage of "hate speech" posts were ascribed to persons of color or how many "misinformation" posts were by members of a given religious affiliation? Do the companies' enforcement actions appear to disproportionately impact vulnerable voices?

Regardless of whether users share their demographics with each report above, companies should be required to provide daily or weekly summaries that list each specific policy and the demographic breakdown (user-reported, purchased from data brokers or inferred by the platform's own algorithms) of enforcement actions taken under that policy. For example, a policy on Covid-19 falsehoods would include a daily table listing by demographic how many enforcement actions were taken against each demographic. The inverse would also be provided, with a table that shows each distinct demographic combination (for combinations with more than X users) and a histogram for that demographic combination of all of the policy violations for that group.

This is critical to understand whether policies are inadvertently disproportionately impacting certain groups such as women or minorities. For example, are hate speech policies inadvertently being enforced more often against women [301] or minorities? [302] Are fact checks being enforced more against certain demographics?

Increased Access to Facebook's Fact-Checking Database

Facebook provides an internal dashboard to fact-checking organizations that lists the posts it believes may be false or misleading. [303] Today, access to that dashboard is extremely limited, but broadening access to policymakers and the academic community as a whole would enable much closer scrutiny of the kinds of material Facebook is focusing on. Given that the company already shares this content with its fact-checking partners, there would be fewer privacy implications to broadening that access to a wider pool of researchers.

Increased Access to Facebook Research Datasets

[300] https://www.npr.org/2019/03/28/707614254/hud-slaps-facebook-with-housing-discrimination-charge
[301] https://www.dailymail.co.uk/news/article-9106809/LEAH-HARDY-gossiping-Alec-Baldwins-Spanish-wife-breached-Facebooks-standards.html
[302] https://www.nytimes.com/2020/10/08/business/black-linkedin.html
[303] https://www.poynter.org/fact-checking/2018/we-asked-19-fact-checkers-what-they-think-of-their-partnership-with-facebook-heres-what-they-told-us/

Through academic partnerships and programs like Social Science One, [304] Facebook permits large-scale research on its 2 billion users, from manipulating their emotions [305] to hyperlink data sets [306] to more in-depth analyses of the flow of information across its platform. [307] Researchers from across the world have been given access to study misinformation and sharing on Facebook, [308] and a closer look at the projects approved to date [309] suggests the kinds of access they have been granted would also support work into understanding the biases of Facebook's own moderation practices.

Offline Harm & The Legal System

Many of the "community guidelines" enforced by social platforms are, at least on paper, also violations of U.S. law, including libel, harassment and threats of violence. How often do social media companies or recipients of those messages refer them to law enforcement and what was the outcome of those cases? If few such posts are ever referred to law enforcement, why do social platforms believe harassment and threats of violence should not be reported to officials if they believe they are dangerous enough to warrant removal from their platforms? Tracking cases where posts were referred to law enforcement and the resulting legal decisions would shed light on how closely social media platforms' interpretations of U.S. laws adhere to reality.

Companies routinely remove content like protest announcements by citing offline harm. A special category of the removal databases above should include moderation actions where companies cited offline harm as the primary reason for removal. This includes any cases where protest calls were removed, since such actions extend the companies' reach into the offline world.

Self-Submission Database Of Private Posts

For private content, social platforms could be required to offer users a one-click button to submit the removed content and explanation from the company to a public database.

Users would be able to share what they believe to be an incorrect removal [310] with the world. High-profile users routinely share such incorrect removals through the media, but this would offer ordinary users the ability to gain visibility for their removals. Forcing social platforms to include one-click submission would also allow researchers and journalists to verify that the removal is real. Certain classes of content like illegal material could be flagged as simply a "PhotoDNA match" or a match into a recognized terrorist content database without further detail or flagged as a non-consensual intimate image, which would still give researchers sufficient information to understand broad patterns.

Similar to the public post database, this should include the full explanation of the takedown and any demographic selectors the user is willing to share.

---

[304] https://socialscience.one/
[305] https://www.pnas.org/content/111/24/8788
[306] https://dataverse.harvard.edu/dataverse/socialscienceone
[307] https://socialscience.one/blog/first-grants-announced-independent-research-social-media%E2%80%99s-impact-democracy
[308] https://items.ssrc.org/from-our-programs/social-media-and-democracy-research-grants-grantees/
[309] https://items.ssrc.org/from-our-programs/social-media-and-democracy-research-grants-grantees/
[310] https://www.dailymail.co.uk/news/article-9106809/LEAH-HARDY-gossiping-Alec-Baldwins-Spanish-wife-breached-Facebooks-standards.html

Algorithms & Interfaces: Engagement Versus Enlightenment

What is the purpose of a social media platform? Is it to generate an endless cycle of mindless engagement to keep users spending as much time as possible viewing ads? Or is it to help society reach a higher level of enlightenment? The design of social platforms today is designed largely to favor the former, using a combination of recommendation algorithms and interface design techniques to keep users glued to their screens.

Should platforms use algorithms and interfaces designed to keep us captivated in a perpetual state of engagement, hand-fed a steady diet of what we want rather than what we individually need to become more informed citizens? Such questions are far from new. In 1961 FCC Chairman Newton Minow famously described television as a "vast wasteland" filled with a "procession of game shows, formula comedies about totally unbelievable families, blood and thunder, mayhem, violence, sadism, murder, western bad men, western good men, private eyes, gangsters, more violence, and cartoons. And endlessly, commercials -- many screaming, cajoling, and offending. And most of all, boredom." [311]

He argued that "it is not enough to cater to the nation's whims; you must also serve the nation's needs" and that advertisers should be "less concerned with costs per thousand and more concerned with understanding per millions." [312] As to whether television should provide the public what it wanted or what it needed, he offered: [313]

> *If parents, teachers, and ministers conducted their responsibilities by following the ratings, children would have a steady diet of ice cream, school holidays, and no Sunday school. What about your responsibilities? Is there no room on television to teach, to inform, to uplift, to stretch, to enlarge the capacities of our children? Is there no room for programs deepening their understanding of children in other lands? Is there no room for a children's news show explaining something to them about the world at their level of understanding? ... There are some fine children's shows, but they are drowned out in the massive doses of cartoons, violence, and more violence.*

Instead, today's algorithmic recommendation systems epitomize this idea of chasing ratings. These systems are designed to keep users spending time on platforms, feeding each person an endless firehose of material custom curated just for them. From radicalizing users into terrorists to spreading viral falsehoods, these silent algorithms are designed with the single purpose of keeping users on their site, producing content to lure in other users and consuming ads. [314] [315] Social media platforms are "designed environments that support particular practices while discouraging others." [316] Engagement-based algorithms "driv[e] views but also privileg[e] incendiary content, setting up a stimulus–response loop that promotes outrage expression" and encourages "polarizing, impulsive, or antagonistic behaviors." [317]

---

[311] https://www.americanrhetoric.com/speeches/newtonminow.htm
[312] https://www.americanrhetoric.com/speeches/newtonminow.htm
[313] https://www.americanrhetoric.com/speeches/newtonminow.htm
[314] https://www.scientificamerican.com/article/youtubes-recommendation-algorithm-has-a-dark-side/
[315] https://firstmonday.org/ojs/index.php/fm/article/view/10419/9404
[316] https://www.nature.com/articles/s41599-020-00550-7
[317] https://www.nature.com/articles/s41599-020-00550-7

Moreover, the specifics of how such algorithms work are closely guarded secrets, meaning society isn't able to have an informed debate about their impact.

Given the centrality of social platforms to today's society, should algorithms be designed to push users towards content hand-selected to interest or enrage them? Or should an algorithm's job be to use what it knows about a person to guide them towards a greater understanding of their world, filling in gaps in their knowledge and nudging them towards intellectual rather than emotional pursuits?

The user interface design of social platforms is designed to make the production and consumption of content as trivial as possible. Even the most miniscule of changes to these interfaces can have an enormous impact on behavior. In the leadup to the 2020 election, Twitter added a small amount of "friction" to retweeting. Users attempting to retweet a post would be presented with a textbox asking them to add their own commentary to the post, in the "hope it will encourage everyone to not only consider why they are amplifying a Tweet, but also increase the likelihood that people add their own thoughts, reactions and perspectives to the conversation." [318] Instead, this small change led to a 20% reduction in retweets, decreasing the total volume of Twitter itself by as much as 70 million tweets a day. [319]

Could changing the algorithms and design of social platforms change online behavior in a form of technological determinism? [320] Or will users simply adapt social platforms to their desires no matter how they are designed? [321]

Legislative Solutions

There is already legislative momentum around restrictions on certain "dark patterns" [322] and "nudge techniques" [323] in which platforms encourage users to "like" and share content as much as possible and use the broadest possible privacy settings. [324] Platforms could also be required to permit users to disable algorithmic recommendation in favor of strict chronological ordering and/or have control over the specific signals used to prioritize content. Most importantly, companies could be required to provide transparency around how their recommendation algorithms work and the rationale for each of their interface design decisions and whether it is designed primarily to increase engagement.

Educational Solutions

[318] https://www.realclearpolitics.com/articles/2021/01/29/tracking_twitters_growth_did_trump_ban_cause_a_dip_145154.html

[319] https://www.realclearpolitics.com/articles/2021/01/29/tracking_twitters_growth_did_trump_ban_cause_a_dip_145154.html

[320] https://en.wikipedia.org/wiki/Technological_determinism

[321] https://en.wikipedia.org/wiki/Social_shaping_of_technology

[322] https://www.wired.com/story/how-to-spot-avoid-dark-patterns/

[323] https://www.theverge.com/2019/8/1/20749517/social-network-legislation-hawley-privacy-research

[324] https://www.bbc.com/news/technology-47933521

Most users of social platforms have little understanding of the myriad ways in which algorithms and design decisions influence what they see. There is an urgent need for K-12 students to better understand that these interfaces are not neutral or benevolent gatekeepers, but rather are designed to suck them in and keep them consuming and producing content.

Technical Solutions

Perhaps the most obvious technical intervention would be increasing the friction around content creation and sharing, as Twitter did prior to the 2020 election. For example, Twitter experimented earlier in 2020 with encouraging users to read an article before sharing a link to it. [325] [326] Allowing users to switch to chronological ordering rather than algorithmic recommendations could also help.

An intriguing idea would be to create an algorithmic version of Newton Minow's vision of encouraging enlightenment over entertainment. This could take a form of algorithmic recommendation that, instead of prioritizing content based on the likelihood that a user will watch or engage with it, would recommend scholarly and informative content that will fill gaps in the user's knowledge and help them become a more informed citizen of the world. Making this the default algorithmic recommendation algorithm could help shape how people interact and engage with social platforms and even turn them into extensions of the educational system.

Balkanization, Echo Chambers & Loss Of Viewpoint Diversity

The concept of "serendipitous discovery" in information science describes the way in which users unexpectedly encounter information of interest or relevance to them. [327] [328] [329] The printed newspaper or linear television broadcast are classic examples in that a user interested in one particular story must thumb through the paper or sit through the broadcast to find what they are looking for, during which time they will encounter may other stories of potential interest to them which they never sought out or knew existed.

In contrast, the web was designed around the concept of non-linear access [330] and information "mobilization." [331] Non-linear access means that rather than reading an entire newspaper, each individual article is broken into its own distinct URL or social media post, isolating users from the rest of the day's news. Mobilization means that instead of having to read an entire article, an individual sentence, quote or photograph can be extracted and circulate by itself, transforming content from fixed publications into collections of ad-hoc individually disseminatable snippets.

---

[325] https://www.theguardian.com/technology/2020/jun/11/twitter-aims-to-limit-people-sharing-articles-they-have-not-read

[326] https://twitter.com/twittersupport/status/1270783537667551233?lang=en

[327] https://www.emerald.com/insight/content/doi/10.1108/00220410310472518/full/html

[328] https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.10359

[329] https://dl.acm.org/doi/abs/10.1145/1640233.1640279

[330] https://books.google.com/books?hl=en&lr=&id=swCAOcQUSn4C&oi=fnd&pg=PA163&dq=mobilization+of+information+hypertext&ots=5DgkygeJfZ&sig=AnbLwbVQL8oyFd_q3fb-XleNjl8#v=onepage&q&f=false

[331] https://www.sciencedirect.com/science/article/pii/S0306457398000612

The ability to access information directly has enabled modern search engines and social media platforms to provide us a precision-curated firehose of individually tailored information. Instead of a cross-section of available information, algorithms curate an online world just for us that adjusts in realtime to a combination of our interests and what the platforms believe will yield the greatest engagement. [332]

The end result is billions of personalized echo chambers, algorithmically curated to make us spend as much time online and to drive an emotional response that encourages us to post a response. [333] Like a child learning how to push its parent's buttons, these algorithms silently watch our every digital action, building exquisitely detailed dossiers of our interests, dreams and fears and the topics and views that drive an emotional response from us in order to feed us ever more of the same.

This balkanization of the national discourse dates back to the nation's founding, with the party paper newspaper system delivering tailored echo chambers based on political affiliation. The professionalization of journalism acknowledged the dangers of such tailored information and pushed news reporting towards the focus on objective all-topics reporting that dominated the latter half of the twentieth century. The rise of cable television in the 1980s prompted a renewed societal debate over the fracturing of the societal discourse, as the big three of ABC, CBS and NBC gave way to a vibrant universe that eventually spanned thousands of interest-specific channels.

Just as social media ads today are precisely targeted based on fine-grained demographics and interests, such targeting came to television news alongside the rise of cable. This "new approach revolutionized the selling of TV shows in the 1980s and enabled networks to keep high-quality shows on the air without catering to the mass audience." [334] Even the newspaper comic "Calvin and Hobbes" commented on this fracturing of the national consciousness when the title character offered in 1992 that "We've got to get cable tv … people across the country are watching different tv shows than we are … if we don't all watch the same tv, what will keep our culture homogeneous?" [335]

Twitter offers a glimpse at just how much of an echo chamber our social platforms are becoming. A decade ago, less than 20% of the tweets sent each day were retweets. [336] Today that number is more than 50%. [337] From less than 1% a decade ago, around 10% of all tweets each day today are either by a verified user or are a retweet of a verified user's post. More than 80% of daily tweets mention another user, [338] but just 30% are replies representing back-and-forth conversation. [339] In short, Twitter is where we go to share

[332] https://www.nature.com/articles/s41599-020-00550-7
[333] https://www.nature.com/articles/s41599-020-00550-7
[334] https://www.washingtonpost.com/archive/lifestyle/tv/1989/12/24/the-80s-were-big-for-tv/fce422b1-9857-4335-a1f6-ecb2461ac8c6/
[335] https://www.gocomics.com/calvinandhobbes/1992/08/29
[336] https://www.forbes.com/sites/kalevleetaru/2019/03/04/visualizing-seven-years-of-twitters-evolution-2012-2018/
[337] https://www.realclearpolitics.com/articles/2021/01/29/tracking_twitters_growth_did_trump_ban_cause_a_dip_145154.html
[338] https://www.realclearpolitics.com/articles/2021/01/29/tracking_twitters_growth_did_trump_ban_cause_a_dip_145154.html
[339] https://www.forbes.com/sites/kalevleetaru/2019/03/04/visualizing-seven-years-of-twitters-evolution-2012-2018/

the thoughts of others, especially elites, rather than our own ideas and to try and get others' attention rather than engaging in mutual dialog.

Even journalists are increasingly retreating into these echo chambers, with NBC's Lester Holt summarizing the growing argument with "I think it's become clearer that fairness is overrated … The idea that we should always give two sides equal weight and merit does not reflect the world we find ourselves in. That the sun sets in the west is a fact. Any contrary view does not deserve our time or attention." [340]

Half a century ago, the growing consolidation of newspapers [341] and the limited broadcast options meant the country shared "a common and generally measured baseline from which to understand and debate the nation's issues. News consumers didn't have media echo chambers in which to retreat. National and international news came from network television or from wire service accounts in local newspapers." [342] This limited media ecosystem created a common reference point for the national debate, even if it was far from representative of the nation's rich diversity.

As Former FCC Chairman Newton Minow recently put the loss of this common reference, "Fractionalization of the audience provides more choice, but we pay a big price. Our country now is much more divided because we do not share the same news or believe the same facts. I used to think providing more choice was in the public interest but I am not sure today." [343]

Legislative Solutions

In 1949, Congress attempted to address similar concerns of echo chambers and the loss of viewpoint diversity by creating the Fairness Doctrine that required the presentation of opposing views. [344] Its elimination in 1987 led directly to the rise of unabashedly partisan personality-driven programming like Rush Limbaugh's show, [345] [346] which were not economically viable under the former model. This suggests that any attempts at legislatively correcting for echo chambers will likely have unintended consequences in actually reducing viewpoint diversity.

Rather than focusing on publishers, Congress could focus on the social platform gatekeepers that increasingly control the public's access to information. Rather than requiring them to meet certain topical metrics, however, like with the Fairness Doctrine, Congress could require that social platforms exempt elected officials and accredited news outlets from content moderation decisions, while enforcing greater transparency on their "recommender" and "trending topics" algorithms to allow unintended or intended political, topical and other biases to be scrutinized by external experts.

[340] https://thehill.com/homenews/media/545803-lester-holt-warns-media-against-giving-a-platform-for-misinformation
[341] https://transition.fcc.gov/osp/inc-report/INoC-1-Newspapers.pdf#page=4
[342] https://thehill.com/opinion/technology/541882-cronkite-signed-off-40-years-ago-it-seems-like-an-eon-in-news-standards
[343] https://www.wsj.com/articles/the-most-popular-shows-youve-never-seen-11613944330
[344] https://en.wikipedia.org/wiki/FCC_fairness_doctrine
[345] https://www.wsj.com/articles/SB122990390599425181
[346] https://www.poynter.org/reporting-editing/2021/how-rush-limbaughs-rise-after-the-gutting-of-the-fairness-doctrine-led-to-todays-highly-partisan-media/

Educational Solutions

As Lester Holt's comments illustrate, American society has reached a point where it no longer sees the value in hearing alternative ideas. Much as the New York Times half a century ago saw the LGBTQ communities as "perverts" [347] and "deviants" [348] to be hunted down and excluded from society, rather than having their voices heard, society today is increasingly viewing those who differ from them as unworthy of a voice in the digital public square.

At the K-12 level, students should be taught the value of reaching beyond their own perspectives to at least understand opposing views and develop the reasoning and conflict resolution skills necessary to conduct informed debates with those with different views, backgrounds, demographics and lived experiences. Emphasizing the value of hearing different perspectives and the dangers of "group think" could help future generations understand why it is so important to be exposed to a diversity of perspectives, even those outside society's mainstream as LGBTQ voices were half a century ago.

Technical Solutions

The single central firehoses of content that define modern social platforms are detrimental to community building, but ideally suited for exposing users to alternative perspectives. At the same time, platforms must be cautious not to inadvertently entrench preexisting views [349] [350] or increase attacks on underrepresented communities by increasing their exposure to aggressive and vocal groups that oppose them.

Instead of attempting to arbitrate "truth" or "acceptable" views or forcing opposing users into contact with one another, a better option might be to merely increase users' access to a greater diversity of information. In areas where users have deeply entrenched views, such additional perspectives may have little impact or simply further entrench existing views, but in areas where users have not yet developed an opinion, such approaches could help mitigate the impact of echo chambers. Even simply seeing quantitatively how much attention different communities are paying to a story could help widen viewpoint diversity. For example, users who argue strongly for and against increased immigration in the United States could find joint interest in seeing that CNN and MSNBC suddenly stopped covering the topic after Vice President Harris was named immigration lead. [351] Or that media coverage of immigration under Donald Trump shifted from people to walls – from human beings to the barriers keeping them out. [352]

---

[347] https://www.nytimes.com/1950/04/19/archives/perverts-called-government-peril-gabrielson-gop-chief-says-they-are.html

[348] https://www.nytimes.com/1950/05/20/archives/inquiry-by-senate-on-perverts-asked-hill-and-wherry-study-hears.html

[349] https://www.forbes.com/sites/kalevleetaru/2017/03/23/the-backfire-effect-and-why-facebooks-fake-news-warning-gets-it-all-wrong/

[350] https://www.forbes.com/sites/kalevleetaru/2018/01/08/facebooks-fake-news-backfire-why-silicon-valley-must-grow-up-from-neverland/

[351] https://www.realclearpolitics.com/video/2021/04/28/immigration_disappears_from_tv_news_after_harris_named_immigration_lead.html

[352] https://www.realclearpolitics.com/articles/2019/02/15/how_immigration_coverage_shifted_from_people_to_barriers_139491.html

Such insights give both sides of a debate insights on how the issues are being covered by the media or policymakers without taking a side and provide opportunities for opposing sides to come together over a common insight.

Another simple example would be to show a basket of several major news outlets, such as CNN, MSNBC and Fox News television channels or a collection of major online outlets and display a graph beside each tweet as to how much coverage that topic is receiving across them. This would help users instantly identify whether coverage of the given topic is highly skewed (suggesting a potentially partisan topic) or evenly distributed. [353] [354]

Conflicting Expectations Of "Netiquette" And Online Aggression

What are the accepted rules for communicating on social media? Is politeness prioritized? Is profanity acceptable? Is sarcasm embraced? Are name calling and personal attacks permissible? Is "doxing" [355] allowable? Are pithy emotional diatribes preferred over calm clinical citations? Should disagreements be resolved through eloquent debate, simply walking away or violently threatening the other person until they leave?

Such questions form the heart of what is acceptable online behavior, also known as "netiquette." [356] They also form one of the most basic forms of conflict on social media over differing expectations and norms of interaction.

Disagreements over how to act online date back to the earliest forms of computer-based communication, accelerating as the digital world became mainstream. [357] [358] Literature of the early to mid-1980s is filled with observations like "widely shared norms on the Usenet are rather few" [359] and "computer-mediated groups exhibited more uninhibited behavior - using strong and inflammatory expressions in interpersonal interactions." [360] The Wall Street Journal once described the Usenet as "the rough-and-tumble saloons and honky-tonks sprawled on the wrong side of the tracks" filled with "oceans of talk, hyperbolic rhetoric, public brawls and damage control" where political newsgroups were "no place for the timid: This is where some of cyberspace's best writers and meanest street-fighters hang out, where incivility is all too common, and no misstep goes unchallenged." [361]

---

[353] https://www.realclearpolitics.com/video/2021/04/26/the_cuomo_stories_have_disappeared_from_television_news.html

[354] https://www.realclearpolitics.com/video/2021/04/30/television_news_coverage_of_the_suez_canal_freighter_versus_the_egyptian_train_crash.html

[355] https://en.wikipedia.org/wiki/Doxing

[356] https://en.wikipedia.org/wiki/Etiquette_in_technology

[357] https://dl.acm.org/doi/10.1145/357431.357435

[358] https://academic.oup.com/jcmc/article/2/4/JCMC2410/4584388

[359] https://files.eric.ed.gov/fulltext/ED334620.pdf

[360] https://www.sciencedirect.com/science/article/abs/pii/0749597886900506

[361] https://www.wsj.com/articles/SB84108829691364000

Considerable early research focused on the complexities of conflict resolution in the digital world, from expressing emotion through text to the unique psychological elements of communicating over distance, with online toxicity emerging as an early challenge: [362]

> *Various constraints operate to prevent excessive verbal aggression in face-to-face or traditional mediated encounters. For example, during a face-to-face encounter, verbal aggression beyond a certain limit will provoke either physical aggression, or intervention by others to prevent further escalation. Both these inhibiting factors - external intervention, and fear of sanction (physical violence) - are absent in bulletin board communication. If the other person is rude, the only sanction one can apply is to be rude in return, and this can escalate due to the absence of any inhibitory constraints. There is very little feeling of an immediate social circle in which people act as self-regulating participants.*

In the Usenet era, this was addressed, like today, primarily through the availability of moderated groups. [363] In moderated groups, a common set of behavioral and topical rules were published for all users to review and each new post was reviewed by an administrator before being published to ensure compliance with these rules, known as "prereview" moderation. In contrast, today's social platforms primarily use "post review" moderation in which content is published publicly and then optionally reviewed at a later date if flagged by a user. One of the reasons platforms are investing so heavily in AI-powered algorithmic moderation is to restore this kind of prereview moderation to ensure violating content is never seen by users.

It is worth noting that as the volume of Usenet posts increased, some moderated groups adopted algorithmic "robo-moderation" to automate the rejection of violating posts, offering a reminder that the current trend today towards automated moderation [364] has a long historical precedent.

Unlike today's social platforms, most Usenet groups were entirely unregulated, with moderated groups explicitly identifying themselves. Each moderated group was free to set its own rules. In many cases two versions of a group would exist, one completely unfiltered and a second with ".moderated" on the end of its name covering the same topic, but with a set of rules governing appropriate behavior enforced by a moderator. Users whose posts were rejected by the moderated group could still publish to the unmoderated version or find another moderated group on the same topic whose policies permitted their posts. In a small number of cases, "excessive verbal aggression on a particular newsgroup … provoked the spawning of a 'nice' version of that newsgroup" [365] that required posters to avoid certain behaviors.

What makes netiquette so important to the future of social platforms is that improving online behaviors could help us eliminate a lot of toxicity from social media that doesn't fall into the classical categories of moderation but can be a major obstacle for underrepresented communities. Flame wars, name calling, threatening language, taunts, overt criticism and the like, even when not targeting a person's protected demographics, can make online communities far less inclusive. Improving conflict resolution online would go a long way towards making it more welcoming for all.

[362] https://files.eric.ed.gov/fulltext/ED334620.pdf
[363] http://pages.swcp.com/~dmckeon/mod-faq.html
[364] https://fortune.com/2020/11/19/facebook-ai-content-problems-artificial-intelligence/
[365] https://files.eric.ed.gov/fulltext/ED334620.pdf

Legislative Solutions

First Amendment protections have long excluded so-called "fighting words" that "by their very utterance, inflict injury or tend to incite an immediate breach of the peace," [366] while permitting most other forms of argumentative speech. However, over the years this concept has been steadily eroded and lost significant meaning in the digital world where communication at a distance means arguments cannot so readily come to blows. This suggests one possible avenue of legislative intervention could be clarifying the point at which an online threat of violence crosses the line from merely discouraged speech to illegal speech prosecutable under the law. [367] Such clarification could also help social platforms and their users know at what point an online threat should be referred to law enforcement, allowing the court system to resolve such disputes rather than the employees of social media companies and ensuring that such speech does not escalate to offline harm.

Educational Solutions

Much of the long-term solution to online aggression and netiquette that is more welcoming to all is likely to come in the form of teaching K-12 students online conflict resolution strategies similar to how educators have long helped shape such behaviors offline on the playground. Teaching students to restrain their emotions when angry, thinking about how their words make others feel online and navigating the more uninhibited nature of "communication at a distance" would help create future generations that are better equipped to navigate the digital world.

Technical Solutions

There is much social platforms could learn from the long history of "computer mediated communication" research into online behaviors and aggression stretching back to the earliest messaging systems more than half a century ago. [368] As evidenced by today's social toxicity, no single system has emerged as a panacea against online aggression, but the design decisions of many previous communications technologies in the web's history offer some intriguing ideas for today's social platforms.

Social platforms today force all users into a single shared communicative space. On Twitter every public tweet is visible to every other user who can mention and reply to them, meaning the most aggressive and rhetorically violent users are able to harass at will. Emotionally charged topics like politics appear alongside mundane discussions of the weather. While Facebook offers the concepts of Groups and selective visibility of posts, by default posts are visible to a user's entire list of friends.

In contrast, Usenet was based around the concept of myriad independent communities, each focused on a specific topic or topics and with its own specific rules on behavior. Users could participate in as many communities as they desired, presenting a different persona in each. On Twitter, an academic interested in scholarly debates about women's rights in the Middle East must contend with every troll and opinionated person that stumbles upon their discussion, with no way to limit their debates to just other scholars in their field. Usenet solved this issue by design, even allowing for subcommunities to split off at will to focus on more narrow subtopics or in response to a disagreement. Moderated groups could limit

[366] https://www.law.cornell.edu/wex/fighting_words
[367] https://via.library.depaul.edu/cgi/viewcontent.cgi?&article=1080&context=jatip
[368] https://en.wikipedia.org/wiki/PLATO_(computer_system)

who participates in debates and excluded users could form their own parallel moderated groups, ensuring both that all voices could be heard by forming their own communities and that communities could establish their own rules that protected their members.

Twitter allows users to manually block other users and attempts to automatically hide some content under a warning message asking "Show additional replies, including those that may contain offensive content?" Facebook's Groups is the closest to the Usenet model, but it still enforces a single persona on users, in which they must use a single user account that ties different sides of their lives together. In contrast, Usenet users could and did use different email accounts to represent different portions of their lives, for example a work email for technical discussions and a personal email for other discussions. The rapid expiration of messages and lack of global search of early Usenet also made it more difficult for harassers to stalk a user across groups.

This raises the question of whether social platforms should do more to replicate that concept of a collection of distinct communities rather than emphasize virality by pushing users to share their posts with the widest possible audience. Twitter notes that "people are allowed to post content, including potentially inflammatory content, as long as they're not violating the Twitter Rules," while acknowledging that "Twitter lets us participate in broad conversations and connect with people from many corners of the world. While hearing from more people can be enriching, it can also be a source of frustration and misunderstanding." [369] Its primarily recommendation for conflict resolution is to just to "block and ignore" offending users, [370] but this fails to acknowledge that the mob mentality of social platforms like Twitter mean that large numbers of users can pile on with abuse.

Following in the footsteps of Usenet's model of unfiltered, moderated and the occasional "nice" groups, what might this model look like for Twitter? One could imagine parallel Twitters, one the current raw unfiltered firehose where users can be as aggressive and confrontational as they like. On the other end might be "nice Twitter" in which only polite clinical discourse is permitted. Users would be free to engage in divisive societal debates, so long as they expressed themselves in clinical language and followed traditional scholarly norms like avoiding profanity and name calling or personal attacks and cited every statement of fact to its source. [371] In between could be myriad "community moderated Twitters" in which a community of users could come together to establish a set of acceptable speech for their version of Twitter that would be enforced only for that version of Twitter. Infinite parallel Twitters could exist, each with its own conduct rules. Users could move between versions of Twitter, finding one that best fits their needs.

Each of these ideas would likely conflict with the economic reality that any segmentation of social platforms' user bases would reduce engagement. Reducing conflicts and confrontations would also likely reduce engagement, by preventing the kinds of pile-ons that accompany viral arguments. As Twitter's brief experiment with adding "friction" to retweeting demonstrates, [372] even the smallest of changes can

[369] https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content

[370] https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content

[371] https://www.forbes.com/sites/kalevleetaru/2019/06/14/could-forcing-thoughtful-civil-discourse-save-social-media-from-its-toxicity/

[372] https://www.realclearpolitics.com/articles/2021/01/29/tracking_twitters_growth_did_trump_ban_cause_a_dip_145154.html

have devastating impacts on their usage, which in turn impacts advertising revenue, suggesting companies may be highly reluctant to explore these ideas.

Constraining Access To Factual Information

The United States has long protected the publication of factual information under most cases, even embarrassing information. A politician who is arrested for accepting a bribe cannot order the media not to cover the story. This is not the case everywhere and even within the United States, support is growing for new legislation that would permit citizens to censor factual information about themselves. [373]

The European Union has the "Right to be Forgotten," [374] under which any citizen may force search engines to remove links to content they deem embarrassing, including factual information. For example, convicted murderers and child abusers can have information about their cases removed from search indexes such that people searching their names will not see anything about their criminal histories. [375] [376] Initially the law was seen as granting EU citizens the right to have embarrassing information removed globally from all countries, but this was later narrowed to just within the EU, [377] though with some exceptions. [378] Similarly, the United Kingdom has the concept of "super-injunctions" [379] that permits citizens to both censor information and prevent anyone from knowing it has been censored.

Today the very technology companies that once fought strongly against the Right to be Forgotten, [380] now offer the equivalent of the UK's super-injunctions to American citizens. For example, when mainstream news outlets reported that one of the co-founders of the Black Lives Matter movement had purchased a portfolio of high-end properties despite publicly endorsing Marxism and raising questions (later dismissed) about whether donated funds had been misallocated, [381] social platforms moved swiftly to ban sharing of links to those stories. [382] [383] As a public figure helping the lead the national debate over changing the American economy, such coverage was highly relevant to the debate over wealth.

Moreover, in keeping with the British super-injunction model, neither company would comment when asked whether they had removed the links at the individual's request. It also raises the question of

[373] https://www.pewresearch.org/fact-tank/2020/01/27/most-americans-support-right-to-have-some-personal-info-removed-from-online-searches/
[374] https://en.wikipedia.org/wiki/Right_to_be_forgotten
[375] https://nakedsecurity.sophos.com/2019/12/02/convicted-murderer-wins-right-to-be-forgotten-case/
[376] https://www.nytimes.com/2019/09/24/technology/europe-google-right-to-be-forgotten.html
[377] https://www.washingtonpost.com/technology/2019/09/24/google-scores-major-victory-eu-right-be-forgotten-case/
[378] https://www.reuters.com/article/us-eu-alphabet-content/facebook-can-be-forced-to-remove-content-worldwide-after-landmark-eu-court-ruling-idUSKBN1WI0QL
[379] https://en.wikipedia.org/wiki/Super-injunctions_in_English_law
[380] https://www.reuters.com/article/us-eu-alphabet-content/facebook-can-be-forced-to-remove-content-worldwide-after-landmark-eu-court-ruling-idUSKBN1WI0QL
[381] https://www.dailymail.co.uk/news/article-9463851/Black-reporter-LOCKED-Twitter-criticizing-BLM-says-company-gone-far.html
[382] https://www.dailymail.co.uk/news/article-9463851/Black-reporter-LOCKED-Twitter-criticizing-BLM-says-company-gone-far.html
[383] https://www.dailymail.co.uk/news/article-9477097/Facebook-blocks-users-sharing-DailyMail-com-story-BLM-founders-property-empire.html

whether elected officials are similarly eligible to have public interest stories about them removed. For example, take the hypothetical example of a presidential candidate whose campaign focuses on replacing capitalism with socialism and prohibiting ownership of more than one home, when multiple news outlets break the story that they actually had a net wealth of hundreds of millions of dollars and a vast property portfolio across the country. Asked whether they would ban all sharing of those stories if the candidate requested, neither Facebook nor Twitter would comment. In 2020, both companies banned sharing of the New York Post's reporting on a laptop allegedly owned by Hunter Biden in the weeks before the election, showing a willingness to censor information that could impact a presidential election. [384]

That governments might use these censorship powers to stifle debate and criticism is far from hypothetical. When farmers in India mounted a mass protest over government policies, Twitter moved quickly to suspend at least 250 prominent users to silence the dissent, only reversing its decision after international outcry. [385] Yet two months later it silenced factual reporting of soaring Covid-19 infections and deaths in the country at the government's request. [386] [387]

In the United States, the companies have closely regulated posts about Covid-19 under dedicated pandemic speech policies. [388] [389] Yet changing scientific understanding demonstrates the dangers of these policies. Today, Facebook's policy prohibits posts that claim "that wearing a face mask does not help prevent the spread of COVID-19." Yet that was precisely the official messaging of US health officials like Anthony Fauci early in the pandemic, [390] even while other countries were recommending mask wearing. Asked whether Facebook would have deleted Dr. Fauci's statements had its Covid-19 policy been in effect at the time, a spokesperson demurred, offering that it is "a good illustration of why clear guidance/rules of the road [via the government] would be helpful. Private companies shouldn't be making these calls on their own, and we've been clear about that."

In mid-March 2021, Facebook's rules prohibited any statements "about the safety or serious side effects of COVID-19 vaccines" or that "COVID-19 vaccines kill or seriously harm people." [391] After EU regulators advised there was a potential link between vaccines and blood clots, [392] Facebook quietly updated its policy to continue banning any mention of blood clots associated with vaccines "except in relation to specific vaccines for which public health authorities have found possible links or are officially investigating such reports." [393]

The ability of governments and private individuals to censor factual, but embarrassing information, runs counter to the long history of speech protections in the United States that has historically protected factual information. As social platforms increasingly permit third party "fact checking" organizations to

[384] https://nypost.com/2020/10/14/facebook-twitter-block-the-post-from-posting/
[385] https://www.bbc.com/news/world-asia-india-56007451
[386] https://www.buzzfeednews.com/article/pranavdixit/twitter-blocking-tweets-india
[387] https://www.medianama.com/2021/04/223-twitter-mp-minister-censor/
[388] https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy
[389] https://www.facebook.com/help/230764881494641
[390] https://www.realclearpolitics.com/articles/2020/04/03/virus_experts_early_statements_belie_prescient_portrayal_142845.html
[391] https://web.archive.org/web/20210315072047/https://www.facebook.com/help/230764881494641/
[392] https://www.ema.europa.eu/en/news/covid-19-vaccine-astrazeneca-benefits-still-outweigh-risks-despite-possible-link-rare-blood-clots
[393] https://web.archive.org/web/20210331005344/https://www.facebook.com/help/230764881494641/

determine "truth" this problem will only continue to grow. [394] In fact, Twitter is now experimenting with allowing ordinary users to decide what they believe is true and false and should and should not be reported on Twitter. [395]

Legislative Solutions

Given the apparent growing interest among the American public for some form of domestic "Right to be Forgotten" legislation, [396] there is considerable need for legislative clarity around whether and under what circumstances American citizens and elected officials should have the right to require social media companies to remove accurate factual information about themselves they find embarrassing. The right of private companies to constrain public access to critical health information during a pandemic, whether by government request to silence criticism or by their own decisions as to what they believe is harmful for the public to see, such as self-reported vaccine side-effects, should also be clarified.

If Congress determines that removal of factual information relating to public interest categories like elected officials, public figures and public health information is not in the best interests of the nation, legislation, including amending Section 230, could clarify under what conditions such removals could or could not occur. Congress could also mandate that internet companies record all removals, including those they themselves perform under their misinformation policies, be recorded to public databases like Lumen [397] to permit independent review of their actions.

Educational Solutions

As internet companies are increasingly willing to remove access to factually correct information, societies must learn to be less trusting of internet gatekeepers, learning strategies to reach around these removals to identify such censored information.

Technical Solutions

The Lumen database [398] is an industry standard repository where internet companies can voluntarily report legal demands to remove content from their platforms, including governmental orders. [399] Social platforms could extend their reporting to include all public-access content they remove or restrict for any reason, lending far greater transparency around their censorship of factual information.

Control Over Government Speech

[394] https://www.realclearpolitics.com/articles/2019/08/24/a_small_number_of_fact-checkers_now_define_our_reality_141087.html
[395] https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html
[396] [396]
[397] https://lumendatabase.org/
[398] https://lumendatabase.org/
[399] https://www.medianama.com/2021/04/223-twitter-mp-minister-censor/

Social media today isn't just where the citizenry of democratic nations gather to debate their collective future. It is where their elected representatives gather to publicly debate, to communicate their policies and beliefs to the public and where they listen to and engage with their constituents. At the federal level, "social media is near-ubiquitous among members of Congress" with most members maintaining both personal and official accounts. [400] In just the first half of 2020, "members of Congress have collectively produced an average of 73,924 tweets and 33,493 Facebook posts each month, generating a total of over 476 million reactions and favorites and over 112 million shares and retweets." [401] Use of social media is expanding rapidly, with Congress producing 81% more tweets and 48% more Facebook posts per month than just four years ago. [402] As Facebook itself notes, social media has become the medium of choice for political communication, "from the President of the United States to your local school board official." [403]

This rapid adoption means that the official speech of government is increasingly coming under the purview of private companies and their acceptable speech policies. In turn, companies are increasingly embracing this role as political censor, deleting posts and threatening to silence lawmakers with whom they disagree [404] and even suspending and banning elected officials, including heads of state.

The early days of radio were marked by similar concerns, with stations banning political candidates and ideas with which they disagreed or felt were dangerous for society. Concern over the ability of private companies to censor the democratic debate led to regulation that heavily constrained the ability of broadcasters to censor candidates for federal office. This included largely barring them from censoring libel or threats of violence by candidates in most cases under the concern that stations could otherwise simply deem any speech they disliked as false or dangerous. In particular, the ability of companies to control the speech of the very politicians that had power over their industries was viewed as antithetical to the concept of a democratically elected government having power over all.

Social platforms today are also increasingly the way citizens engage with their elected officials to have their voices heard. While constituents can still mail letters, make phone calls and send emails, those are typically read and responded to by staff, whereas on social media, any citizen in the nation can speak unfiltered directly to the President of the United States and receive a response back in realtime. This means that when a user is banned from social media, they lose this direct conduit to their elected representatives, placing them at a disadvantage to their fellow citizens. In essence, the conduct of government business and campaigning over social media implicitly assumes that all citizens have a chance to participate, whereas in reality social platforms decide which citizens should have the right to participate.

This represents something fundamentally new in the history of communication technologies, in that while past technologies like the post office were filtered, they did not comprise such a fundamental core of the political discourse. Similarly, while newspaper commentaries are one way for citizens to influence the public debate, they do not represent the same kind of back-and-forth conversation of social platforms. Moreover, the plethora of newspapers in the United States means even if one outlet refuses to run a commentary, another may, whereas social platforms tend to largely enforce similar speech policies.

[400] https://www.pewresearch.org/internet/2020/07/16/1-the-congressional-social-media-landscape/
[401] https://www.pewresearch.org/internet/2020/07/16/1-the-congressional-social-media-landscape/
[402] https://www.pewresearch.org/internet/2020/07/16/1-the-congressional-social-media-landscape/
[403] https://www.engadget.com/facebook-oversight-board-rules-for-politicians-donald-trump-211353831.html
[404] https://www.foxnews.com/opinion/twitter-tried-censor-me-they-lost-sen-tom-cotton

Legislative Solutions

The history of broadcast regulation and the decades of court battles that shaped and defined their role over political speech offers a template for similar protections that could be carved out of Section 230. Under the broadcast standard, social platforms likely could no longer apply fact checking warnings directly to official government posts, delete posts or suspend or ban elected officials. It would also have likely prevented Facebook from deleting Elizabeth Warren's campaign ads calling for its breakup. [405]

In contrast, the broadcast standard would still have likely permitted social platforms to suspend Donald Trump under the provisions of its imminent harm exception, though the courts would likely have to resolve how such rules applied to official government policy statements.

As government increasingly conducts its business over social platforms, there is a need for further clarity as to just what right citizens should have, if any, to have their voices heard on those platforms and what rights policymakers have to protect their policy statements from deletion.

Educational Solutions

There is an urgent need for greater public and policymaker awareness of the control social platforms wield over the official statements of government in order to spark a broader societal debate over whether additional protections are needed.

Technical Solutions

Given that many social platforms already treat elected officials as a special category of user and often have separate enforcement guidelines for their posts, [406] the technical capabilities are already in place to exempt their speech from moderation. In terms of ordinary users, one solution would be to permit users who are banned from platforms for any reason to continue to be able to engage with elected official accounts to ensure their voices are still heard. In essence, instead of being deactivated, they would be transitioned to a special limited account able only to communicate with elected officials of their government.

Debating In Realtime

Social platforms enable us to experience the world in realtime, without respect to geography. We can watch an event happening on the other side of the world moment by moment through the eyes of those participating in and witnessing it, while simultaneously hearing the reaction from across the world and participating in the global discussion ourselves. The ability to experience events from far away through local eyes in realtime is often described as a fundamentally new capability created by social media, but it was actually radio which first introduced this concept to the public.

[405] https://www.politico.com/story/2019/03/11/facebook-removes-elizabeth-warren-ads-1216757
[406] https://help.twitter.com/en/rules-and-policies/public-interest

As Michael Stamm puts it, with the advent of radio "American homes became filled with the sounds of news, speeches, music, and church services, all broadcast live from what were to listeners unseen and often distant sites." [407] Television brought the ability to see those events as well as hear them. As then-FCC Chairman Newton Minow put it, Alan Shepard's pioneering spaceflight was "witnessed by millions of anxious Americans who saw in it an intimacy which they could achieve through no other medium, in no other way." [408]

The difference is that these previous mediums merely broadcast voices from afar into people's homes, while providing no way for listeners to respond back. This made it very different from in-person debates. As a 1934 editorial put it: [409]

> *political spell-binders, who sought to help themselves by blacking their rival's character, learned that such efforts at a street corner rally would instantly be met by a challenge by answering hecklers, or by persons who rose to make utterly serious, truthful denials of the defamation just spoken against a fellowman. At least, it took nerve to face this risk. The radio speaker, on the other hand, stands in a well-guarded room with all of civilized society, laws and police force to protect him from any interruption, no matter how false or uncivilized his attacks on other human beings may be.*

In contrast, social media replicates the traditional face-to-face environment of debate, extending it to a global audience. Today when the Chairman of the Federal Reserve speaks, listeners engage in lively debates in the live commentary sections that run alongside the livestreams. [410]

Usenet first introduced the general public to this kind of live interactive societal debate, with the Wall Street Journal once likening the roiling raging political debates of the day to the "the sort of family screaming match sometimes found at booze-soaked Thanksgiving dinner." [411] Yet Usenet lacked social media's emphasis on realtime reaction – when engaging in a debate, "there's no need to wake up in the middle of the night with the snappy comeback you wish you'd had available -- you can spend the whole night, if you wish, crafting a devastating response." [412]

In contrast, social media prioritizes realtime reaction over all else. Users no longer have the luxury of spending days carefully researching and writing a response - they are expected to respond instantly. As fact checkers have long lamented, by the time the facts are known about a breaking news story, most users have moved on to the next story, leaving their outdated and wrong responses as digital detritus.

The realtime nature of today's social debates can also encourage mob mentality, in which users pile into a debate, rushing to join the fray, rather than stepping back and considering all of the facts and perspectives. Algorithmic recommendation systems tend to surface "trending topics" which in turn drives ever more users into the heated debate of the moment.

---

[407] https://books.google.com/books?hl=en&lr=&id=PXHuUO_UJi4C
[408] https://www.americanrhetoric.com/speeches/newtonminow.htm
[409] https://scholarship.law.nd.edu/cgi/viewcontent.cgi?article=4082&context=ndlr
[410] https://www.wsj.com/articles/when-fed-chief-talks-so-do-listenersand-they-provide-an-earful-11618604698
[411] https://www.wsj.com/articles/SB84108829691364000
[412] https://www.wsj.com/articles/SB84108829691364000

Most importantly, realtime conversation tends to emphasize the raw visceral emotional reactions of stream-of-consciousness discourse. Upon encountering something that upsets us online, we are expected to immediately respond and register our outrage, feeding the outrage cycle.

Legislative Solutions

The challenges of debating in realtime do not readily lend themselves to legislative intervention, though clarifying the legal landscape around libel and "fighting words" online could lend clarity to the legal exposure created by such debates.

Educational Solutions

Education at the K-12 level could focus on online conflict resolution and constructive debating strategies, along with helping students understand the potential legal ramifications of libel and threatening behavior online could help mitigate some of the negative behaviors associated with realtime debating.

Technical Solutions

For more than three decades major US stock exchanges have implemented so-called "circuit breakers" that are "designed to slow trading down for a few minutes, to give investors the ability to understand what's happening in the market, consume the information and make decisions based on market conditions." [413] Applied to social platforms, such a circuit breaker could flag any pair or group of users who are rapidly replying to one another with an escalating level of emotion in their posts and gradually slow their messages down, eventually pausing them from posting for a few minutes in a "cool down" period. Relying on the "velocity" of a conversation (the rate of back-and-forth messages) would help avoid the issues of simply warning about confrontational language use. [414]

Recommendation algorithms could also be redesigned to consider the emotion of trending topics. Topics with largely clinical or observational language (such as first-hand reports of a breaking event) would be recommended as before, as could those with high levels of non-aggressive emotion (allowing for both positive stories and mourning of negative events like deaths) but topics with a surge in aggressive language would be ineligible for trending recommendations. This would prevent algorithmic "pile-on" mob debates.

Everyone Is A Publisher & Every Voice Is Equal

In 1993 the New Yorker published an iconic cartoon of a dog using a computer, proclaiming to a fellow dog that "On the internet, nobody knows you're a dog." [415] [416] Even in its infancy, there was a "wariness

[413] https://en.wikipedia.org/wiki/Trading_curb
[414] https://www.theguardian.com/technology/2020/jun/11/twitter-aims-to-limit-people-sharing-articles-they-have-not-read
[415] https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog
[416] https://www.washingtonpost.com/blogs/comic-riffs/post/nobody-knows-youre-a-dog-as-iconic-internet-cartoon-turns-20-creator-peter-steiner-knows-the-joke-rings-as-relevant-as-ever/2013/07/31/73372600-f98d-11e2-8e84-c56731a202fb_blog.html

about the facile façade that could be thrown up by anyone with a rudimentary knowledge" of webpage creation. [417] Indeed, in 2016 a random website titled "WTOE 5 News" looked official enough to help seed the international story that the Pope had endorsed Donald Trump for president. [418] Today no coding knowledge is required to share one's voice with the world - just a smartphone and an internet connection.

Yet this great power of the internet is also its greatest weakness: it made everyone with internet access an international publisher.

In the latter Usenet era, the email addresses from which users posted offered a form of identity, context and gatekeeping. Often, a person's email address offered clues to their real-world identity. Addresses associated with major university, commercial or governmental institutions conferred a level of gatekeeping trust. A post to a scientific newsgroup about a new materials science discovery might carry more weight coming from an email address associated with a major university materials science professor than from a random person on the internet with a long history of posting false conspiracies to that newsgroup. Today, even email no longer conveys this trust, as academics, journalists, business leaders and even government officials routinely email using the same @gmail.com or similar email provider email address as any other person, making it impossible to know if they are who they claim to be. [419]

On social media these critical contextual cues of identity and association are absent by design. On Twitter, users are free to select any username and profile image they please. This is profoundly empowering for members of underrepresented communities in that they can choose how to present themselves to the world. Those who have experienced discrimination because of their race, gender or other characteristics can choose not to present those attributes to the world in order to be evaluated purely on their ideas. This selective anonymity also empowers activists and whistleblowers to communicate more openly and freely than they could if their posts were associated with their real identities.

At the same time, it means that social media lacks any form of trusted context through which to evaluate the information we consume. A post from a senior official at the CDC speaking in the official capacity of the United States Government to announce a new policy has no more authority or credibility on Twitter than a mischievous teenager in their basement, a scammer or an adversarial foreign government looking to sow confusion. The CDC official might have a blue check mark (meaning a "verified" account) beside their name confirming they are the person they claim to be, but such verification only confirms that the name on the account is accurate, it makes no representation beyond that. Moreover, given that users must explicitly request a verified account, it is more likely that a celebrity influencer will have a blue check mark beside their name than a CDC scientist, lending them more "credibility" in Twitter's user interface than the career CDC scientist.

The Covid-19 pandemic reminds us of the dangers of a digital world in which all voices are equal. The inability of the public to easily distinguish between authoritative voices speaking on behalf of the government, scientists deeply involved in the response, expert commentators with deep understanding, criminal scammers, foreign misinformation actors, mischief makers and well-meaning but misinformed

[417] https://www.washingtonpost.com/blogs/comic-riffs/post/nobody-knows-youre-a-dog-as-iconic-internet-cartoon-turns-20-creator-peter-steiner-knows-the-joke-rings-as-relevant-as-ever/2013/07/31/73372600-f98d-11e2-8e84-c56731a202fb_blog.html

[418] https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html

[419] https://www.politico.com/news/2021/04/09/lego-gamer-infiltrated-white-house-press-480673

citizens creates the perfect environment for dangerous information chaos. All of these voices are equal on social media.

Worse, the algorithmic curation of social platforms tends to prioritize voices that yield the highest engagement. This means the government official with a handful of followers speaking in calm clinical and highly technical wording is less likely to go viral than the incendiary falsehood-laden post by the misinformed celebrity that is carefully timed and promoted to spread as rapidly as possible.

This suggests that a critical path towards mitigating online misinformation lies in restoring some form of identity context to social media, allowing users to understand a bit about a poster, their expertise and potential motivations.

This need for greater context and trust around real-world identity extends beyond social media to online resellers. For example, when purchasing a product in a bricks-and-mortar store, there are established legal requirements and precedents that mean customers can for the most part trust that what they buy from the shelf isn't banned in the US over safety concerns. Online, there are no such guarantees as companies double down on allowing third parties to sell on their websites with little direct oversight. Shoppers on Amazon might assume that the company carefully vets each and every product for sale on its website, ensures their authenticity and safety and lists only reputable sellers. The reality can be far different, despite the company blocking more than three billion suspicious listings in 2018 alone. [420]

Legislative Solutions

One area that would likely receive bipartisan support would be clarifying whether e-commerce sites are responsible for ensuring that the products they sell are authentic and are not known to be unsafe under guidelines similar those of physical stores. [421] A first step towards this would be requiring e-commerce sites to make more information about sellers available to customers, including the company's legal operating name and details on what other storefronts on the site are operated by the same seller under different names.

Increased regulation around sponsored content on social media would make it easier for users to understand the motivations of a poster, while requiring platforms to make it easier to see all content sponsored by the same company would make it easier to understand the context around a post.

Educational Solutions

Teaching K-12 students how to think critically about the motivations of a given post and how to research the identity of the person or organization behind it would go a long way towards helping future generations navigate the online world. For example, upon encountering a social media post advocating for a particular position, students should recognize that they should first research the poster before

[420] https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990

[421] https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990

examining its arguments and should have the skills to research who the poster is and determine whether, for example, it is a registered lobbying or advocacy firm.

Technical Solutions

Most social platforms today offer the concept of a "verified user" in which the platform certifies that the owner of the account is the person it publicly claims to be. However, this is typically limited to verification of identity, not expertise. One possible solution would be to include a "verified degrees and affiliations" section in which a user could list their formal degrees and certifications and their employer and role there, along with other formal positions such as volunteer organizations. Thus, a person claiming to hold an MD from Harvard Medical School and a computer science PhD from Stanford and be employed by the CDC and serve on several government advisory boards could have this information confirmed by Twitter and displayed in a special section of their profile. Users seeing posts from them about the CDC's work on computational modeling of Covid-19 could then trust that this person is likely speaking more authoritatively about that work than someone without those credentials and affiliations.

At the same time, adding formal degrees and affiliations to social platform identities, while restoring trust, also strips away the equalizing power at the heart of social media: that an ordinary citizen has just as much voice as the president. How might platforms establish trust for activists who wish to remain anonymous and ordinary citizens with powerful ideas but without formal degrees or relevant employment?

One possibility would be an icon beside each Twitter user's account that displays a popup with a "you are here" map of where they are situated in the Twitterverse. This would show a sample of the accounts they follow and that follow them, who retweets them and whom they retweet, what news outlets have cited their tweets and who posts similar kinds of content with similar arguments. Thus, a human rights activist who is heavily retweeted and followed by prominent human rights organizations and leaders and whose posts routinely appear in major press can establish a degree of authority without revealing their actual identity. In essence, such a visual would create a social platform version of the citation networks that underlie academic trust, in which the more an author is cited in their field, the more trust may be ascribed to their work.

Such a display would also help restore some of the institutional memory that governed Usenet newsgroups. In the Usenet era, a user who posted daily uninvited conspiracy theories to an unmoderated newsgroup would become known to members and either ignored or have their posts treated with a high level of skepticism. On Twitter, the lack of segmentation into distinct communities means that a conspiracy theorist who posts on a wide range of topics can encounter new users each day who don't know their background and history. Changing the interfaces of social platforms to emphasize posters' previous activity and history, from the people they engage with to the kinds of things they post, would help restore some of this institutional memory.

Monopoly Power Over Communication

Why does it matter what Twitter or Facebook's speech rules are? After all, if you don't like one platform, you can simply use another or build your own. The problem is that a handful of platforms have become so dominant they wield effective monopoly power over social media and collaborate to enforce largely identical rules on acceptable speech.

When Uber banned conservative commentator Laura Loomer for her tweets, Lyft moved in lockstep, [422] effectively barring her from 98% of the ridesharing market in the United States. [423] When Airbnb bans a user for hate speech, [424] they are cut off from a platform that accounts for 20% of domestic lodging expenditures and growing. [425] When Amazon refuses to publish a book that violates its acceptable speech policies, [426] the author is cut off from the source of 72% of new adult book sales. [427]

Moreover, like the voluntary cooperative censorship agreements of the motion picture and broadcasting eras, companies today tend to move in unison with censorship decisions. When one company bans a person or class of speech, its peers move in lockstep. As Twitter and Facebook banned Donald Trump, companies including Apple, Discord, Google, Pinterest, Reddit, Shopify, Snapchat, Stripe, TikTok, Twilio, Twitch and YouTube all joined in banning Trump or related content within short order. [428]

The fact that Twitter and Facebook competitors Snapchat and foreign-owned TikTok all moved to ban Trump alongside their peers serves as a stark reminder that in today's digital world, Twitter and Facebook largely set the rules for everyone, even foreign companies.

Similar industry-wide guidelines were seen in the early era of motion pictures and in broadcast to this day, but newspapers have always offered a bulwark against their narrowing views with a greater diversity of perspectives, back to the party papers of America's founding days. No matter how many news outlets refuse to cover a story, some outlet somewhere can cover it. Yet in a world in which 86% of Americans consume news digitally today and more than half turn to social media [429] and two-thirds to search engines to find that news, [430] the speech rules of social platforms are increasingly shaping even the news we see. [431]

Even the core infrastructure of the web is increasingly centralized into the hands of just a few companies whom increasingly view their services not as the neutral plumbing of the internet but rather as an extension of their own speech policies. When Twitter competitor Parler grew to number one on Apple's App Store by permitting many of the topics and individuals banned by its peers, it was silenced from the web [432] and mobile devices [433] for violating their speech policies.

---

[422] https://www.nbcnews.com/news/us-news/laura-loomer-banned-uber-lyft-after-anti-muslim-tweetstorm-n816911

[423] https://www.vox.com/2018/12/12/18134882/lyft-uber-ride-car-market-share

[424] [424]

[425] https://www.vox.com/2019/3/25/18276296/airbnb-hotels-hilton-marriott-us-spending

[426] https://abigailshrier.substack.com/p/book-banning-in-an-age-of-amazon

[427] https://www.wsj.com/articles/they-own-the-system-amazon-rewrites-book-industry-by-turning-into-a-publisher-11547655267

[428] https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html

[429] https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/

[430] https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

[431] https://nypost.com/2020/12/10/how-media-covered-up-the-hunter-biden-story-until-after-the-election/

[432] https://www.buzzfeednews.com/article/johnpaczkowski/amazon-parler-aws

[433] https://www.theverge.com/2021/1/9/22221730/apple-removes-suspends-bans-parler-app-store

Legislative Solutions

It is critical to recognize that the absence of corporate monopolies is not the same as an absence of speech monopolies. When Twitter and Facebook banned Donald Trump, they were joined by companies from across industries, including foreign-owned competitors. The long history of industry-wide speech codes from the motion picture and broadcasting eras demonstrate how entire industries of competitors can join together to create common speech rules enforced by all. Thus, anti-trust legislation in its traditional sense is unlikely to significantly change the landscape of monopoly speech rules.

The experience of Parler reminds us of the immense power wielded over the internet by those who control its central infrastructure. One potential policy intervention would be to bar companies who account for more than a certain percentage of online infrastructure (including cloud computing, mobile device access, ISPs, etc) from barring a company from using their services because of their speech. A customer could still be banned for technical or legal reasons like computer hacking and violations of US laws like copyright infringement, but they could not banish a company like Parler purely because they disagree with its speech policies.

Similar protections could prevent companies that account for a dominant portion of their respective industry from barring a user over their speech, along with providing special protections for accredited news organizations. Thus, Uber and Lyft could no longer bar a user for their speech, while social platforms would be barred from refusing to link to news outlet websites just because they disagree with a story.

Educational Solutions

There are few obvious educational pathways to addressing Silicon Valley's consolidation over the online world, though one option would be to teach students about the history of centralized speech control both in the US and globally to help inform their understanding of the ramifications of social consolidation today.

Technical Solutions

There have been a number of initiatives over the past several years like the Data Transfer Project [434] to make it easier for users to download all of the data they have contributed to a given social platform over time and be able to upload that to a competing social platform. [435] The problem is that as social platforms increasingly move in lockstep on their speech policies and enforcement actions, simply moving to another platform doesn't change anything. As Parler's experience attests, competitors that stray from the party line to offer a more expansive list of acceptable speech rules are simply removed from the internet.

This suggests that while data portability and interoperability are important, the challenges surrounding the increasingly centralized and coordinated speech policies of the web cannot be solved through technical means alone.

[434] https://datatransferproject.dev/

[435] https://www.eff.org/deeplinks/2018/07/facing-facebook-data-portability-and-interoperability-are-anti-monopoly-medicine

No Representation In Rule Making

One of the greatest sources of conflict over social platforms today is that their users have little say over their policies. Senior executives or their founders make unilateral decisions that determine the online rights of billions of users. [436] When Twitter banned Donald Trump, the decision was made by its chief legal counsel with Jack Dorsey's direct approval. [437] After Twitter banned him, Zuckerberg personally approved the removal of two of his posts and then banned him. [438] Decisions as consequential as banning the president of the United States are not made by democratic boards working with outside advisers and holding a vote. They are made by an individual billionaire making a personal judgement call based on what they view as best for their company's shareholders.

It was not always this way. In Facebook's early days it was actually a quasi-democracy in which select major policy decisions were put to a vote of its users, with their decisions being binding on the company. [439] The votes were even certified by an outside auditor, [440] with the results of its inaugural Facebook Site Governance vote being reported on April 24, 2009. However, the company selected which policies were put to a vote, ensuring that controversial decisions disliked by users were not available to them to overturn.

As Facebook passed 900 million users it discontinued its experiment with democracy, announcing that "our growing relationship with regulators around the world has created a new layer of accountability with respect to our practices and policies" and thus users no longer needed a say because "we have entered into a settlement agreement with the Federal Trade Commission which involves regular audits of our privacy practices; we work closely with the Irish Data Protection Commissioner's Office, which completed a comprehensive audit of our data practices last year; and we are now subject to the regulatory authority of the U.S. Securities and Exchange Commission." [441] In short, Facebook argued that its users no longer needed a say in its governance, since they could now do so through their governments' elected officials, yet as Trump's experience reminds us, those governments have little say over Facebook's speech policies.

Even Facebook's "Supreme Court," known formally as its Oversight Board, [442] offers no representation to Facebook's 2.6 billion users [443] beyond the ability to submit written comments that can be read by the board. In the United States, the members of the United States Supreme Court are nominated and approved entirely by democratically elected public officials and are all American citizens. In contrast, Facebook's Oversight Board has absolute jurisdiction over Facebook speech policies in every country it is used, despite the Board's combined membership having "lived in" just 27 countries and speaking just 29 languages and including former elected heads of state. [444] If the citizens of a country feel Facebook and

[436] https://www.washingtonpost.com/technology/2021/01/16/how-twitter-banned-trump/
[437] https://www.washingtonpost.com/technology/2021/01/16/how-twitter-banned-trump/
[438] https://www.nytimes.com/2021/01/07/technology/facebook-trump-ban.html
[439] https://www.forbes.com/sites/kalevleetaru/2019/04/13/facebook-was-a-democracy-2009-2012-but-we-didnt-vote-so-it-turned-into-a-dictatorship/
[440] https://www.facebook.com/notes/facebook/results-of-the-inaugural-facebook-site-governance-vote/79146552130/
[441] https://newsroom.fb.com/news/2012/06/the-facebook-site-governance-vote/
[442] https://oversightboard.com/
[443] https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/#:~:text=With%20more%20than%202.6%20billion,most%20popular%20social%20media%20worldwide.
[444] https://about.fb.com/news/2020/05/welcoming-the-oversight-board/

its Board are ruling repeatedly to silence their speech or to permit hate speech against them where such speech has led to genocide, [445] they have no ability to include a representative of their country on the Board.

In fact, the speech of democratically elected US policymakers is now subject to the decisions of former elected heads of foreign governments, meaning the US government no longer has sovereignty over its official speech on one of the most dominant social platforms.

At the same time, merely putting all of Facebook's acceptable speech rules to a global vote once a year would likely lead to underrepresented voices being silenced and wild swings in policy decisions, much as US elections can experience substantial shifts from election to election. In the United States, the Constitution and existence of the court system help provide a stable foundation that protects unpopular views and ensures the stability of government by protecting a basic set of guaranteed protections. A strictly popular vote election would preference larger countries over smaller ones and without the minimum protections of a constitution would permit larger countries to entirely silence the dissenting voices of smaller ones.

Legislative Solutions

Today all levels of government rely on social media to communicate policy decisions and hear from their constituents "from the President of the United States to your local school board official" [446] and nearly the entirety of Congress relies on it. [447] This means that the policy decisions governing who can talk about what on social media in turn governs which democratically elected officials and their constituents are granted a voice in the public debates shaping our nation and which ideas are permitted to participate in those debates. Legislative clarity is required around the transparency of these decisions and whether the users of platforms should have the right to help define their policies or define their oversight boards. Such changes might require amending Section 230 to require companies to provide greater policy representation to their users, though this also raises critical questions around how such representation would work globally and how conflicting policy demands across countries would be resolved.

Educational Solutions

The public has largely accepted the idea that private companies and their billionaire founders should be allowed to define the marketplace of ideas for our democracy and decide on their own what voices and ideas to permit and which to exclude without any representation from the public. [448] Educating the public about the implications of this model and explaining the lack of representation, accountability and transparency in current speech rules could help fuel a broader public debate about whether changes are needed.

Technical Solutions

[445] https://www.oversightboard.com/decision/FB-I2T6526K
[446] https://www.engadget.com/facebook-oversight-board-rules-for-politicians-donald-trump-211353831.html
[447] https://www.pewresearch.org/internet/2020/07/16/1-the-congressional-social-media-landscape/
[448] https://www.reuters.com/article/us-usa-election-trump-social/senior-u-s-democrat-urges-twitter-facebook-to-ban-trump-from-platforms-idUSKBN29C022

The extremely low turnout of Facebook's early voting system [449] offers a reminder that merely offering the technical means to vote does not necessarily translate to an engaged user base interested in expressing their voices. At the same time, additional technical transparency around the creation and implementation of policies, such as internal deliberations and the rationale behind key enforcement decisions would help inform the public debate over the need for public representation in policymaking.

Speech Is Forever On The Web

The Internet's vast decentralized nature means it is nearly impossible to truly delete something completely. No matter how many copies are taken down, other copies can persist on websites in countries with less restrictive speech rules or in darker corners of the web. Online library Sci-Hub's ability to remain available over nearly a decade despite massive copyright infringement reminds us just how impossible it is to truly remove anything. [450]

The eternalness of the digital world means that perpetrators of crimes are increasingly live-streaming video or live-posting imagery of their crimes to social media to ensure a permanent record is saved for posterity. [451] Even if the originals are taken down, downloaded copies can be reposted across the web indefinitely. In fact, Facebook notes that it allows such videos to be shared on its platform so long as it is done to condemn the violence, despite revictimization. Today it is almost routine for imagery of high-profile murder victims to be widely shared across social media, causing perpetual pain for family members and loved ones. [452] [453]

At the same time, the archival nature of social media platforms that preserve a user's tweets through time means that a post as a young adult or even as a child can lead to significant consequences decades later. The once-private developmental period where children and young adults learn about acceptable speech is now part of their permanent public record. It is becoming increasingly common for social media posts from high school [454] [455] or college [456] [457] to resurface years later with grave consequences. Even posts sent as an adult in the heat of the moment [458] [459] or before the person entered the public eye [460] routinely

[449] https://www.forbes.com/sites/kalevleetaru/2019/04/13/facebook-was-a-democracy-2009-2012-but-we-didnt-vote-so-it-turned-into-a-dictatorship/
[450] https://en.wikipedia.org/wiki/Sci-Hub
[451] https://www.nbcnews.com/news/us-news/four-arrested-facebook-live-torture-video-now-charged-hate-crimes-n703456
[452] https://people.com/crime/bianca-devins-stepmom-begs-public-stop-sharing-photos-murder/
[453] https://people.com/crime/at-sentencing-family-of-slain-influencer-confronts-man-who-murdered-her-posted-photos-of-killing/
[454] https://www.nbcnews.com/news/us-news/trea-turner-sean-newcomb-apologize-years-old-racially-insensitive-tweets-n895701
[455] https://www.washingtonpost.com/nation/2019/09/25/carson-king-viral-busch-light-star-old-iowa-reporter-tweets/
[456] https://thehill.com/opinion/civil-rights/544052-forcing-black-teen-vogue-editor-to-resign-over-teen-tweets-is-cancel?rl=1
[457] https://www.nbcnews.com/tech/tech-news/how-delete-old-tweets-they-come-back-haunt-you-n896546
[458] https://www.bbc.com/news/world-us-canada-45052534
[459] https://www.sfchronicle.com/local-politics/article/SFUSD-school-board-member-criticized-for-racist-16039069.php
[460] https://www.bbc.com/news/world-us-canada-43936042

resurface, while mundane historical posts can give away sensitive information about a person's life. [461] So-called self-destructing posts that automatically delete after viewing or after a certain number of hours can still be saved as screen captures and republished to the web.

All of this data is used by companies to construct cradle-to-the-grave biographical dossiers on their users. Facebook's algorithms claim to be able to predict who are interested in treason against their government, [462] who are secretly closeted LGBTQ, [463] who privately hold conservative views,[464] and myriad other sensitive attributes, even if the users have gone to great lengths to conceal this information.

Asked in 2018 whether Facebook would consider removing its predictions of a user's sexual orientation in countries where being LGBTQ is punishable by death, the company refused, arguing that it was important for advertisers to be able to reach them. [465] Asked whether the company had ever received a legal demand from a country to provide a list of its citizens whom the company's algorithms had predicted to belong to a given sensitive category, including ones that could be punishable by death, the company confirmed that it would do so if required under the laws of that country. [466] In short, repressive regimes today can outsource their intelligence collection to Facebook.

Moreover, as more and more of our lives are lived in the digital world, everything from our grocery purchases to our medical records are increasingly being hacked and released on the web, meaning the kind of public scrutiny once associated only with celebrities now befalls even the most mundane members of society.

Legislative Solutions

Possible legal solutions include requiring companies to offer the victims of crimes or their surviving family members the right to have content added to their content signature removal databases that prevent that material from being shared on their platforms. While there has been voluntary momentum around allowing victims of non-consensually shared intimate imagery to have the material blocked, [467] legislation requiring platforms to provide such tools could likely find bipartisan support.

Most social platforms prohibit the sharing of materials derived from computer hacking, which Twitter initially cited when it banned linking to the New York Post's story about the Hunter Biden laptop [468] before changing its explanation. Yet such policies, if evenly enforced, would prevent the publication of stories

[461] https://www.wired.com/story/twitter-location-data-gps-privacy/
[462] https://www.theguardian.com/technology/2018/jul/11/facebook-labels-russian-users-as-interested-in-treason
[463] https://www.forbes.com/sites/kalevleetaru/2018/07/20/facebook-as-the-ultimate-government-surveillance-tool/
[464] [464]
[465] https://www.forbes.com/sites/kalevleetaru/2018/07/20/facebook-as-the-ultimate-government-surveillance-tool/
[466] https://www.forbes.com/sites/kalevleetaru/2018/07/20/facebook-as-the-ultimate-government-surveillance-tool/
[467] https://www.forbes.com/sites/kalevleetaru/2017/11/29/what-facebooks-latest-revenge-porn-effort-gets-wrong/
[468] https://www.realclearpolitics.com/articles/2020/10/16/twitter_facebook__hunter_biden_big_tech_as_big_brother_144467.html

like The Panama Papers [469] or myriad other leaks of information about public officials. Legislation could help clarify that social platforms should ban leaks involving private citizens, but permit leaks pertaining to elected officials or other public figures.

In the European Union, the General Data Protection Regulation (GDPR) [470] was designed to curtail the kind of mass biographical profiling performed by social platforms. However, the reality of its enforcement [471] [472] [473] [474] means its impact has been more limited, reinforcing the myriad workarounds social platforms have found to privacy legislation.

Educational Solutions

In addition to traditional discussions of online dangers, K-12 curriculums should help students understand the permanency of what they say online and how to balance expressing themselves with the fact that that expression today now becomes a part of their permanent societal record.

Technical Solutions

All major social media platforms already implement signature-based content removal technology in the form of PhotoDNA for child exploitation imagery [475] and most use similar tools to detect copyright violations [476] and previously identified terrorist content. [477] Thus, the infrastructure to prevent the sharing of content like non-consensual intimate imagery, crime victim imagery and the like is already widely deployed and such content could simply be added to their databases.

The Globalization Of Local

We live in an era in which social platforms have made it possible for an ever-growing percentage of the earth's population to have their voices heard on the international stage, with the lowliest citizen having the same potential reach as the most powerful head of state. Yet this newly empowered digital citizenry spends their daily digital lives inhabiting a landscape of echo chambers hand-fed by algorithms sifting through the firehose of worldwide commentary to find the posts that will maximize their outrage. [478] An errant comment somewhere on planet earth can now, through the power of these algorithmic echo chambers, become the next viral source of global outrage in countries everywhere as it is taken out of context or reinterpreted to fit myriad agendas all across the world.

---

[469] https://www.icij.org/investigations/panama-papers/
[470] https://gdpr-info.eu/
[471] https://www.forbes.com/sites/kalevleetaru/2019/04/15/us-companies-can-still-harvest-and-resell-eu-citizen-data-under-gdpr/
[472] https://www.forbes.com/sites/kalevleetaru/2019/05/23/gdprs-massive-research-exemption-facebook-and-social-science-one/
[473] https://www.forbes.com/sites/kalevleetaru/2018/12/14/facebooks-latest-breach-illustrates-the-limits-of-gdpr/
[474] https://www.forbes.com/sites/kalevleetaru/2019/05/06/as-gdpr-turns-one-is-it-a-success-or-a-failure/
[475] https://www.microsoft.com/en-us/photodna
[476] https://support.google.com/youtube/answer/3244015?hl=en
[477] https://about.fb.com/news/2017/06/how-we-counter-terrorism/
[478] https://www.nature.com/articles/s41599-020-00550-7

The end result is the globalization of local, in which what were once local discussions limited to small communities are now global debates, with people all over the world weighing in. Anything anywhere can now become the next must-comment event in an instant. In turn, the kind of reactions to every event once required only of elected officials is now mandatory of every organization, whose representatives are expected to weigh in on Twitter on every breaking story.

In September 2020 Coinbase's founder Brian Armstrong pushed back on this growing trend when he published his vision for a "mission focused company" that is "laser focused" on cryptocurrency and would not "engage here when issues are unrelated to our core mission" or "advocate for any particular [political] causes or candidates internally that are unrelated to our mission because … even if we all agree something is a problem, we may not all agree on the solution." [479] The press and punditry outcry was immediate, with a general consensus that in the social media era companies can no longer be "apolitical" or "just opt out" of social activism. [480] Twitter's founder Jack Dorsey condemned Coinbase's announcement, [481] while its former CEO Dick Costolo called it "the abdication of leadership … tech companies used to welcome lively debate about ideas and society. It was part of the social contract inside the company and it's what differentiated tech culture from, say Wells Fargo culture … good luck getting the best engineers in the world to work [there]." [482]

Yet as companies are expected to weigh in on the events of the moment, where is the line drawn between brave companies speaking out on their principles [483] and monopolies "powerful enough to heckle senators with snotty tweets?" [484] As CEOs engage directly with their critics on Twitter, is it acceptable for them to make crude sexual demands of government regulators [485] or allege to the world that a private citizen is a "pedo guy" and "child rapist" as a joke? [486] Should the US military respond to public questioning with "get right before you get left, boomer," [487] or trillion-dollar companies spar with Congress in sarcastic tweets like "you don't really believe the peeing in bottles thing, do you?" [488]

This globalization of local, coupled with the realtime nature of social media, means that the court of public opinion can render a verdict and globally vilify a person based purely on rumor before the subject even has a chance to correct the record. In April 2021, when a Jeopardy contestant used his fingers to show he was a three-time winner, just as he had as a one and two-time winner, within hours he was labeled a white supremacist across social media, [489] with more than 500 previous Jeopardy contestants signing an open letter condemning him [490] and searches for his name now bring up endless pages of social media

[479] https://blog.coinbase.com/coinbase-is-a-mission-focused-company-af882df8804
[480] https://www.fastcompany.com/90562207/can-a-company-really-be-apolitical-in-2020
[481] https://twitter.com/jack/status/1311423420274372608?s=20
[482] https://twitter.com/dickc/status/1311022395491196928?ref_src=twsrc%5Etfw
[483] https://www.wsj.com/articles/ceos-plan-new-push-on-voting-legislation-11618161134
[484] https://twitter.com/SenWarren/status/1375283617341968385?ref_src=twsrc%5Etfw
[485] https://mashable.com/article/elon-musk-sec-taunt/
[486] https://www.bbc.com/news/world-us-canada-50695593
[487] https://www.newsweek.com/marines-apologize-will-adjust-fire-twitter-war-words-tucker-carlson-women-military-1575991
[488] https://twitter.com/amazonnews/status/1374911222361956359
[489] https://nypost.com/2021/04/28/jeopardy-players-demand-apology-over-alleged-white-power-symbol/
[490] https://medium.com/@j.contestants.letter/letter-from-former-jeopardy-2eda854efdf1

posts and news coverage accusing him of displaying a white supremacist sign. [491] [492] [493] [494] This despite fact checking organizations confirming that the hand gesture was simply indicating his number of wins. [495] In the pre-social era, confusion over a hand gesture would have been resolved by media outlets contacting the contestant, asking him to explain his gesture and speaking with independent experts. Instead, the realtime speed of social media and need for the entire world to weigh in on every story meant that a few seconds of an innocent hand gesture became a story of white supremacy and racism.

The globalization of once-local stories can also lead to a lifetime of pain for victims of crimes and their families. The 2019 murder of 17-year-old Bianca Devins [496] would, in the pre-social media era, have been just another tragic story in the local newspaper. Instead, through the power of social media, the images of her life and death have become a global story, republished and repurposed to harass her family and friends and celebrate the heinous nature of her murder. [497] In short, once everyone is a publisher, murderers can ensure the permanent victimization of their victims and their families, while the schoolyard bullies whose reign previously extended as far as the local neighborhood can now reign terror globally, as algorithms pluck formerly local stories for them to amplify and distort.

Legislative Solutions

One area of urgent legislative intervention is the ability for victims of crimes and their families to be able to restrict the circulation of imagery and video that revictimizes them, including the sharing of non-consensual intimate imagery (so-called "revenge porn"). Potential laws that have been proposed include "Bianca's Law" [498] that would require large social platforms to remove "violent and gory content" that violates their policies and the SHIELD Act of 2021 that would bar revenge porn. [499] [500]

While Section 230 provides social platforms with what amounts to near-absolute immunity from publishing libel, important questions remain unsolved as to whether platforms should be required to do more to counter libel. Without narrowing the protections of Section 230, legislators could require that platforms implement "circuit breakers" that would automatically slow down and reduce the visibility of commentary that makes criminal or other damaging allegations against a non-public figure until it is reported on by the news media or alleged in a court filing. This would largely mirror Wikipedia's "Biographies Of Living Persons" policy that requires such allegations to be sourced to a reputable external source rather than to original reporting by the poster. [501]

---

[491] https://deadline.com/2021/04/jeopardy-kelly-donohue-contestant-condemns-white-supremacy-accusations-more-than-i-could-bear-1234746594/
[492] https://news.yahoo.com/jeopardy-contestant-breaks-silence-insists-183536760.html
[493] https://www.cbs8.com/article/entertainment/entertainment-tonight/jeopardy-contestants-call-on-show-to-address-recent-winners-alleged-white-power-on-air-hand-gesture/603-d1573b55-167e-420b-90a5-a60362010846
[494] https://variety.com/2021/tv/news/jeopardy-kelly-donohue-white-power-hand-sign-apology-1234963169/
[495] https://www.snopes.com/fact-check/jeopardy-hand-gesture/
[496] https://en.wikipedia.org/wiki/Murder_of_Bianca_Devins
[497] https://www.syracuse.com/crime/2021/03/brandon-clark-sentenced-bianca-devins-mom-talks-about-indescribable-pain-of-teens-murder.html
[498] https://www.congress.gov/bill/116th-congress/house-bill/8323/text?r=10&s=1
[499] https://amendments-rules.house.gov/amendments/SPEIER_034_xml210311224936258.pdf
[500] https://www.theverge.com/2021/4/15/22340260/vawa-shield-act-revenge-porn-first-amendment-questions
[501] https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons

Educational Solutions

As a society we have largely embraced the idea that everything happening worldwide each day should be available for users from anywhere in the world to comment upon. Indeed, this ideal of empowering every person with a voice lies at the very heart of the social revolution. At the same time, the loss of the concept of "local" means every story is now a global one, with significant implications for privacy, revictimization and online bullying. Societies must be educated in the etiquette of such spaces, from how companies and their CEOs should engage with critics to helping K-12 students learn how to research the context and background of a story before contributing to its spread and recognizing and avoiding harassment and bullying.

Technical Solutions

Social platforms already automatically reduce the visibility and shareability of posts their fact checking partners deem false, meaning they have the infrastructure to detect the core topical focus of a post and slow its spread. [502] Similarly, in February 2021, Facebook began reducing the visibility of "political" content across its platform in the US, showing its ability and willingness to voluntarily deemphasize entire categories of content. [503] [504]

Building on this preexisting capability and willingness, social platforms could build an automatic "circuit breaker" for posts that make allegations of criminal or other damaging behavior against a non-public-figure. Such posts could not be widely shared and would have their visibility reduced until those claims are repeated in a mainstream publication. Extended to public figures this would prevent conspiracy theories like the 2017 story of bodies in barrels on the Clinton's property [505] or Hillary's March 2021 arrest by Navy Seals, [506] from ever going viral in the first place.

Extended to aggressive speech, this could also sharply reduce the prevalence of online toxicity by permitting unfettered debate on a topic while requiring that debate to be in the form of clinical language rather than profanity-laden diatribes and threats of violence.

The Loss Of Community

Comparing early social platforms like mailing lists and Usenet with today's social media platforms, perhaps the most obvious and central difference is the loss of community. Early systems like Usenet were typically based around the concept of a rich landscape of independent communities, each with their own rules, norms and users. [507] In contrast, today's platforms force everyone into a single global community, with every user and every topic forced to share the same common space, whether through Twitter's single

[502] https://www.facebook.com/communitystandards/false_news
[503] https://www.washingtonpost.com/business/2021/02/10/facebook-political-content/
[504] https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/
[505] https://www.politifact.com/factchecks/2017/jun/26/freedomcrossroadsus/saga-bodies-found-barrels-clinton-property-fake-ne/
[506] https://www.reuters.com/article/factcheck-clinton-arrest-seals/fact-check-hillary-clinton-was-not-arrested-by-navy-seals-on-march-2-idUSL1N2L71CF
[507] https://www.wsj.com/articles/SB84108829691364000

firehose of posts or Facebook's shared newsfeed of all one's friends and recommended public page posts from across the world.

Forcing the entire world together into a single space maximizes the amount of content accessible to all users and encourages engagement by forcing users into contact with material they would not otherwise have encountered. At the same time, it maximizes the potential for conflict and hateful and aggressive conduct as vulnerable communities are unable to escape communities that have historically harmed them.

In real life, no city would host the Democratic and Republication national conventions on the same evening on side-by-side stages in the same hotel ballroom – the potential for physical conflict would be too great. Yet, every day on Twitter represents this exact collision, with centrist through extreme members of every partisan persuasion forced into the same common public square to shout over one another. Online toxicity thrives in such environments.

Facebook offers the concept of "Groups" [508] which partially replicate this idea of community, but differ in that users have a single identity (their Facebook user account) to participate in all of the groups they are members of. In Usenet one could maintain multiple personas to participate in different groups by using different email addresses. In contrast, on Facebook, all participation across all groups is tied back to a user's single identity. Indeed, Facebook actually prohibits having multiple accounts. [509]

Facebook's Head Of Product Policy Monika Bickert noted the difficulties inherent in the loss of local community: "We have a really diverse global community and people are going to have very different ideas about what is OK to share. No matter where you draw the line there are always going to be some grey areas. For instance, the line between satire and humour and inappropriate content is sometimes very grey. It is very difficult to decide whether some things belong on the site or not." [510]

Legislative Solutions

The loss of community in online social platforms does not readily present itself to legislative solutions, though one possible area would be to clarify whether users should have the right to maintain multiple independent accounts on social platforms to allow them to separate portions of their lives or minimize the ability of harassers or stalkers to follow them across social platforms.

Educational Solutions

One area of possible educational intervention would be to teach K-12 students how to engage in thoughtful debate in this unified digital public square. Navigating the unique communicative complexities of the online world requires students to understand the implications of this loss of community and that the strategies they use to communicate in person where they are surrounded by their local community are different from a world in which everyone is connected to everyone, including those who wish to hate and harm them.

[508] https://www.facebook.com/help/1629740080681586
[509] https://www.facebook.com/help/203498356357867
[510] https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Technical Solutions

What might social platforms like Twitter and Facebook look like if they restored the concept of "community?" Rather than a single firehose of tweets, Twitter could be reimagined as many independent Twitters, each focused on a specific topic. Rather than tweeting to the world, users signing into Twitter for the first time would be directed to a master listing of all topics and select those they wish to participate in, with a mixture of unfiltered public, moderated public and private groups available. Unlike the Twitter "communities" formed by the use of hashtags, which are still visible to all users, this segmentation of Twitter would restore users' control over whom they encounter online, allowing them to create welcoming spaces, each of which could enforce its own rules over acceptable speech attuned to the needs of that specific community.

Similarly, social platforms could permit users to have an unlimited number of distinct user accounts to use to reflect different parts of their lives, allowing them to partition their digital participation.

The Need For Information Literacy

Beneath the spread of all "fake news," misinformation, disinformation, digital falsehoods and foreign influence lies society's failure to teach its citizenry information literacy: how to think critically about the deluge of information that confronts them in our modern digital age. Instead, society has prioritized speed over accuracy, sharing over reading, commenting over understanding. Children are taught to regurgitate what others tell them and to rely on digital assistants to curate the world rather than learn to navigate the informational landscape on their own. Schools no longer teach source triangulation, conflict arbitration, separating fact from opinion, citation chaining, conducting research or even the basic concept of verification and validation. In short, we've stopped teaching society how to think about information, leaving our citizenry adrift in the digital wilderness increasingly saturated with falsehoods without so much as a compass or map to help them find their way to safety. Instead, we have taught them to blindly trust Silicon Valley to arbitrate "truth" through algorithms, content moderators and crowdsourcing. [511] [512]

Rather than invest in information literacy, society has doubled down on technological solutions to combating digital falsehoods, leaving Silicon Valley to harness legions of "fact checkers," blacklists, content moderators, algorithms and other quick fixes that have done little to turn the tide. The problem is that technology can only mitigate the symptoms, it cannot address the underlying cause of digital falsehoods: our susceptibility to blindly believing what we read on the web and our failure to verify and validate information before we share or act upon it.

When a random website on the web called "WTOE 5 News" proclaimed that the Pope had endorsed Donald Trump, it quickly racked up nearly a million Facebook engagements, [513] with few stopping to click on the site's "About" page that clearly stated it was "a fantasy news website. Most articles on

[511] https://www.forbes.com/sites/kalevleetaru/2016/12/11/how-data-and-information-literacy-could-end-fake-news/

[512] https://www.forbes.com/sites/kalevleetaru/2019/07/07/a-reminder-that-fake-news-is-an-information-literacy-problem-not-a-technology-problem/

[513] https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html

wtoe5news.com are satire or pure fantasy." [514] A handful of Twitter accounts popping up proclaiming, without any evidence, to be government employees "resisting" a new president they disliked were widely embraced by the academic and scientific communities [515] [516] and began fundraising [517] with much of the reaction from the press and scholarly communities being that of open embrace rather than verification and skepticism. [518] Today even society's younger generations that have grown up in the digital world do little better at discerning the credibility of information they see online. [519]

A growing number of countries are emphasizing technical literacy, teaching K-12 students how to code, but technical literacy is not the same as information literacy. The ability to write computer code is unrelated to the ability to perform research, understand sourcing and navigate and resolve conflicting information. To combat online falsehoods, the velocity and virality of social platforms must be replaced with verification and validation.

Addressing the challenge of online misinformation requires recognizing it as a societal challenge requiring education, rather than a technical problem that can be magically solved through code.

Legislative Solutions

Information literacy is largely an educational concern, though legislation could help with funding for curriculum development and training for information literacy programs.

Educational Solutions

Improving the information literacy of America's citizenry cannot be solved by code alone. It requires an in-depth investment in K-12 education to teach students how to think critically, understand primary, secondary and tertiary sources, [520] how to resolve conflicting information under uncertainty and verify and validate everything they see. The disciplines of information science and history are particularly well-suited for teaching these kinds of skills. [521]

The reference desks of public libraries could also play a crucial role in helping their local communities seek out reputable information and combat falsehoods. It is a curious artifact of the modern age that we as a society have these incredible personalized resources in our local communities all across the nation staffed by our next-door neighbors who know us by name, yet increasingly when we are in need, we place our

[514] https://web.stanford.edu/~gentzkow/research/fakenews.pdf
[515] https://www.forbes.com/sites/kalevleetaru/2017/01/25/what-the-rogue-epa-nps-and-nasa-twitter-accounts-teach-us-about-the-future-of-social/
[516] https://www.forbes.com/sites/kalevleetaru/2017/01/28/how-the-rogue-twitter-accounts-rewrote-how-we-communicate-science-in-the-social-era/
[517] https://www.forbes.com/sites/kalevleetaru/2017/01/30/the-rogue-wars-when-the-rogue-twitter-accounts-start-fighting-and-fundraising/
[518] https://www.forbes.com/sites/kalevleetaru/2017/01/25/what-the-rogue-epa-nps-and-nasa-twitter-accounts-teach-us-about-the-future-of-social/
[519] https://ed.stanford.edu/news/stanford-researchers-find-students-have-trouble-judging-credibility-information-online
[520] https://www.crk.umn.edu/library/primary-secondary-and-tertiary-sources
[521] https://www.forbes.com/sites/kalevleetaru/2019/08/05/computer-science-could-learn-a-lot-from-library-and-information-science/

trust in anonymous strangers halfway across the world. To those generations born in the digital age, libraries are often dismissed as outdated museums to a past era, warehouses that rent physical books and DVDs. In reality, libraries are about information and the people and practices to help society navigate the world around them – the perfect solution to today's problems. Perhaps the answer to the deluge of digital falsehoods, fraud and foreign influence lies in a return to our nation's public libraries and their reference librarians that in 2017 alone answered more than a quarter-billion questions for their local communities. [522]

Technical Solutions

Over the last few years, algorithms and armies of fact checkers have been seen as the primary solution to online misinformation, but they cannot replace an information literate society. Using human fact checkers to rate the "truth" of posts and then flag ones deemed false can lead to the "backfire effect" in which suppressed content actually spreads more widely than it would have otherwise. [523] [524] Instead, technical solutions should focus on providing additional context to help users evaluate content they encounter online.

In the Usenet era, the combination of out-of-realtime responding, easy search and threaded responding (in which each comment is posted as a response to the one before it, providing a complete chain of provenance back to the first post) created an environment that was far more conducive to information literacy. As the Wall Street Journal put it in 1996: [525]

> *Posters challenge each other's facts and figures, returning with barrages of data dug up from government documents, well-thumbed reference books, and on-line news archives. Because of newsgroups' archival nature, those who distort questions or twist words come in for particular attack -- questions of who said what are approached with an exactitude that's almost monastic. Writers tend to copy the messages they replying to, pasting them into a new message and juxtaposing their counterpoints with their opponent's points. When executed well, the form lends itself to elegant arguments, advanced with surgical efficiency.*

Could aspects of this environment be recreated and even augmented in the modern era for social platforms? One technically simple solution that would go a long way towards helping contextualize information online would be to restore the provenance that dominated the threaded era of Usenet. When encountering an original post on Twitter today, there is no way to know if that person came up with that post themselves or merely copy-pasted it from somewhere else. Similarly, an image posted to social media must be taken at the face value of the caption the poster assigned to it.

Instead, it would be relatively straightforward for Twitter to display beside each post longer than a few words an icon indicating whether it is unique in the recent history of Twitter and if not, the user can click the icon to see an abbreviated chronology of that message through time, from the first person to post it to Twitter to the present and a timeline showing its ebbs and flows across the platform. Similar

[522] https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey
[523] https://www.forbes.com/sites/kalevleetaru/2017/03/23/the-backfire-effect-and-why-facebooks-fake-news-warning-gets-it-all-wrong/
[524] https://www.forbes.com/sites/kalevleetaru/2018/01/08/facebooks-fake-news-backfire-why-silicon-valley-must-grow-up-from-neverland/
[525] https://www.wsj.com/articles/SB84108829691364000

functionality could be built for public posts on other platforms. Fuzzy matching would ensure that minor edits to a post wouldn't exclude it from being found.

Even such a trivial solution would go a long way towards identifying inorganic campaigns in which large numbers of users are asked to post a certain message in unison. It would also make it easier to root out false first-person reports. A post offering a graphic description of an event the user claims to have just seen minutes before could be tied back to a post from days prior, showing the person had merely copied the earlier post.

Similarly, any image or video posted to social media could undergo a reverse image search that would scan across all public posts on the platform and across the open web for other copies. An image posted to Twitter ten minutes ago and captioned "Live protest in Tehran now" could be instantly identified as a long-circulating image that was taken 10 years ago in Saudi Arabia. Thus, any image shared on Twitter would be accompanied by an icon that links to other copies of it across Twitter and the web and other common captions for it, perhaps with a warning if the description in the current tweet deviates substantially from consensus descriptions.

Connecting imagery and videos to their original sources would allow users to readily see if they have been edited in a misleading way. For example, splicing together unrelated comments or slowing down the video speed to make the speaker seem disoriented or inebriated. [526] Rather than merely affixing a label warning users that a video has been edited (which they may or may not believe), provide one-click access to the original source material so they can see themselves. In short, rather than simply inform users of misleading information, which may simply backfire, convince them through guiding them through the underlying evidence until they themselves arrive at that conclusion on their own.

Such automatic contextualization could also extend to claims of fact made in posts. Social platforms could provide a "citations" capability for posts where users could cite their sources without those citations counting towards the word or character limit of that platform. Thus, a tweet presenting an argument on gun violence could cite the statistics it repeats by linking to the original government sources without exceeding Twitter's character count. Much as it tracks down the originals of images and videos, it would also be possible for Twitter to automatically contextualize statistics by searching news coverage, academic publications and government reports for the given statistic and automatically providing a citation for it for users who wish to dive deeper.

Another intriguing idea would be to offer users a "you are here" map of where a post sits in the Twitterverse. An icon beneath any tweet would display a popup with a network diagram of related topics [527] or tweets discussing the same argument from various sides, along with the names of the posters, grouped together by similarity or how often the users retweet each other. This would provide an at-a-glance view of how the given tweet's arguments are situated in the broader conversation about that topic, giving the user a contextual view to help make their own decisions. This is especially useful in unsettled debates where no clear answer or consensus has emerged.

[526] https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI

[527] https://link.springer.com/article/10.1007/s42001-020-00086-5