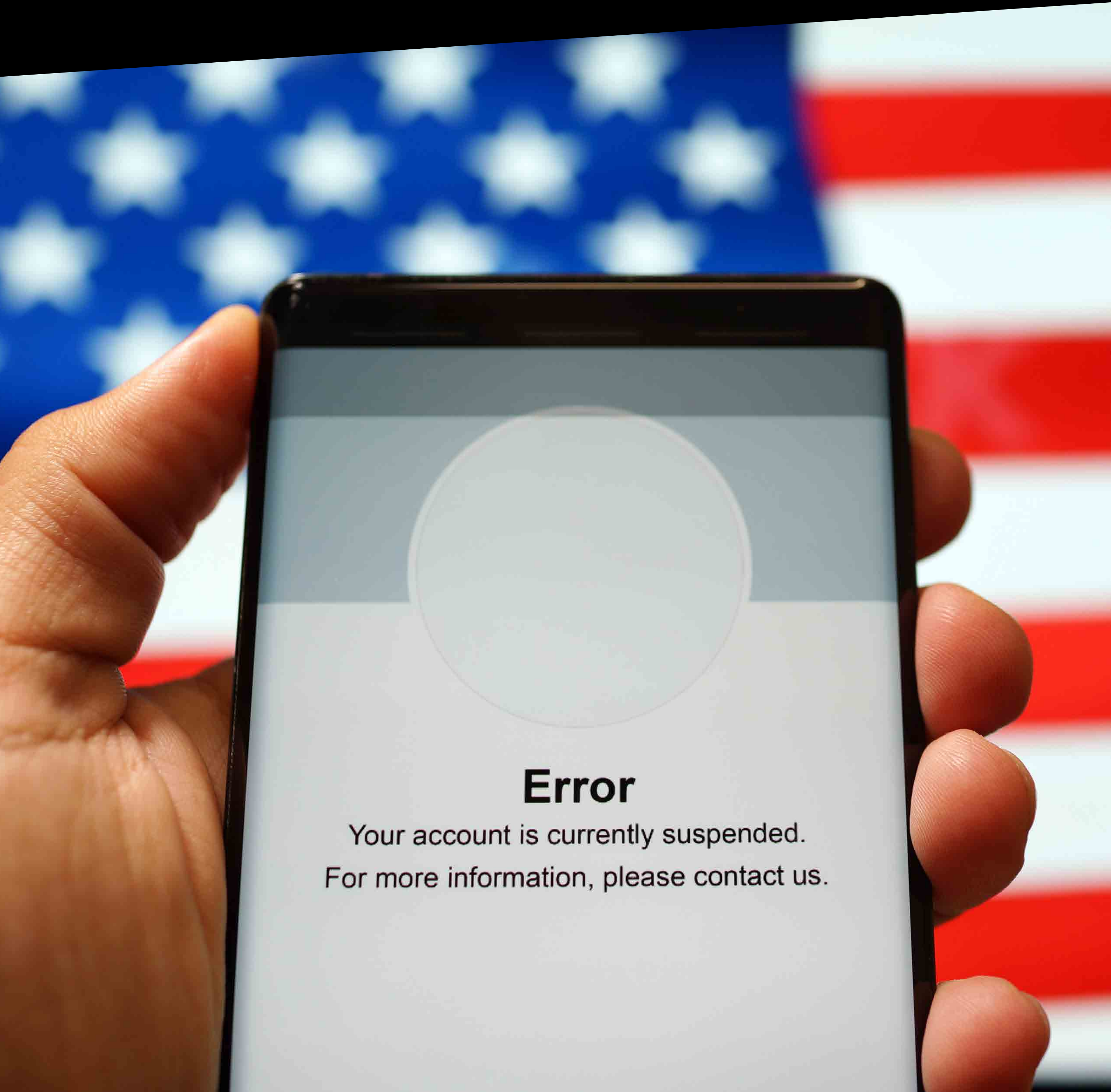
OCTOBER | 2021

RealClear Foundation

Transparency Is the First Step Toward Addressing Social Media Censorship

Kalev Leetaru



Error

Your account is currently suspended.
For more information, please contact us.

To view the full-length research report behind this analysis, including citations, please visit:

https://assets.realclear.com/files/2021/10/1892_leetaru-social-media-digital-censorship-and-the-future-of-democracy-working-paper.pdf

ABOUT THE AUTHOR

Kalev Leetaru—One of Foreign Policy Magazine's Top 100 Global Thinkers of 2013, Kalev is a Media Fellow at the RealClearFoundation. From 2013-2014 he was the Yahoo! Fellow in Residence of International Values, Communications Technology & the Global Internet at Georgetown University's Edmund A. Walsh School of Foreign Service, where he was also an Adjunct Assistant Professor. From 2014-2015 he was a Council Member of the World Economic Forum's Global Agenda Council on the Future of Government. Kalev has been an invited speaker throughout the world, from the United Nations to the Library of Congress, Harvard to Stanford, Sydney to Singapore, while his work has appeared in the presses of more than 100 nations and from Nature to the New York Times. In 2011, The Economist selected his Culturomics 2.0 study as one of just five science discoveries deemed the most significant developments of 2011, while the following year HPCWire awarded him the Editor's Choice Award for Edge HPC (High Performance Computing) "representing the highest level of honor and recognition given to the thought leaders in the HPC community" and in 2013 noted "his research helped usher in the era of petascale humanities." Kalev received a BS in Computer Science and a PhD in Library and Information Science, both from the University of Illinois, where he held the Irwin, Boyd Rayward, Josie Houchens and University Fellowships at the University of Illinois Graduate School of Library and Information Science.

Contact the author at **kleetaru@realclearfoundation.com**

RealClearFoundation is a nonprofit organization working in partnership with RealClearPolitics to mount journalistic and educational programs in the public interest.

EXECUTIVE SUMMARY:

Social media is the communications fabric that increasingly underlies modern society, undergirding democracy itself. It is becoming the public square through which we have our societal debates and the medium through which we speak to our elected officials and they speak back to us. More and more, it is the channel through which governments from local to national publish laws, policies, and regulations to the public, how schools announce schedules, and how companies announce products. It is where we talk to the world and where we talk to one another. Yet within social platforms' walled gardens, society and government are subordinate to private censorship, with social media companies, through their content moderation policies, now deciding what we see and say and even what policies our elected officials are permitted to publicly embrace on their platforms. The censorship rules that social platforms devise thus shape our lives and the future of our nation in unprecedented ways. The rules they devise become, in many ways, the rules of our national conversations about the future of America.

All nations will inevitably confront disagreements over the ideas, narratives, and beliefs that govern them. In democratic societies, the free flow of information and freedom of speech empower the citizenry to discuss and debate and, through the ballot box, to reach a consensus. When that free flow of information and speech is curtailed, with private companies able to decide what is acceptable to see and say, especially when the rules they establish are opaque and their enforcement uneven, the representativeness and legitimacy of those public debates can be corroded.

The growing privatization of the public square in the hands of just a few social media platforms has raised bipartisan interest, with competing demands for increased censorship, decreased censorship, or the creation of alternative platforms. None of the three is likely to solve the underlying issues:

- Democrats, to the extent that they see a problem with social media censorship, generally believe that social media companies should increase their censorship efforts. America's history of broadcast regulation suggests that social platforms will eventually use such censorship to stifle criticism, thwart attacks on their oligopoly status, and advance government policies that benefit their commercial interests over the public good.
- Republicans generally want to strip social media companies of their Section 230 protections, which shield platforms from liability for their content. This proposal would likely result in more censorship, not less, and further entrench the social media oligopoly.
- Libertarians and free-market advocates generally claim that censorship opponents should create their own social media alternatives. Yet, as demonstrated by Parler's experience, Silicon Valley requires that all social platforms enforce the same basic speech guidelines or they are removed from app stores and essentially deleted from the internet.

Before we can debate the future of social platform control over speech, we must have a greater understanding of how well it works today. We must replace today's opaque, subjective, ever-changing, uneven, and unaccountable content moderation with clear, objective, and standardized rules, evenly applied to all. We must also replace today's hands-off approach of trusting social platforms to decide what is best for society with external visibility into the impact that their rules are having.

In short, we need transparency.

Requiring social media companies to fully publish all their policies, guidelines, and precedents, eliminate their unpublished exemptions, clearly explain every decision in plain language, and offer rapid appeals would make moderation more objective and standardized. By removing the veil in front of removal decisions, they would no longer seem as politically motivated or capricious.

Transparency would also likely reduce the total amount of removals by allowing users to better understand what is allowed and disallowed before they post and permit public debate about the merit of those rules. It would reduce error and encourage uniformity by forcing moderators to clearly consider each removal and cite supporting evidence and precedent. It would also lend credibility to those decisions if users had a clear explanation of why their post was viewed as a violation and would make it difficult for companies to maintain parallel rules for politically sensitive users and communities, ensuring that all users are treated equally.

To make this transparency vision a reality, external auditing and societal scrutiny of moderation activities are needed. Social platforms should be required to provide the following ten data sets, which would shed critical light on their operations and enable, for the first time, an open societal debate about their growing influence over the public square:

1. Public algorithmic trending data sets
2. Automatic database of public post violations
3. Database of deleted and exempted protest posts
4. Database of fact-checked posts
5. Database of private post violations by journalists and politicians
6. Demographic database of content removals
7. Increased access to Facebook's fact-checking database
8. Increased access to Facebook's research data sets
9. Database of posts referred to the legal system or removed because of offline harm
10. Self-submission database of private posts

As we consider what a more transparent social platform landscape might look like, a useful model is that of Wikipedia, which has clear rules, a chronologically documented history of all actions, and public archived "Talk" pages where contributors and administrators discuss, debate, and reach consensus, creating trust and enabling external scrutiny and debate over its activities. There is much to Wikipedia's model of transparency that could be adopted by social platforms to lend greater visibility and transparency to their moderation activities.

Ultimately, transparency would shift content moderation from informing users to convincing them and enabling public debate.

INTRODUCTION:

PRESERVING THE FOUNDATION OF SOCIAL MEDIA REQUIRES TRANSPARENCY

The World Wide Web was seen in its earliest days as nothing short of the greatest democratizing force ever created, while the rise of social media was prophesied to give voice to the disenfranchised, underrepresented, and silenced voices of the world. Twitter once touted itself as “the free speech wing of the free speech party” and even rebuked Congress’s initial calls for it to ban terrorists, citing that absolute free speech trumped all other considerations. Over the years, this utopian dream has given way to an emphasis on “healthy conversation” and ever-changing rules defining “acceptable speech.” 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 such as “What do we do when a movement is authentic … but is inherently harmful?” 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.

For most of their existence, social media platforms largely avoided censoring elected officials and mainstream news outlets in the United States. That all changed over the last few years as Silicon Valley began labeling President Trump’s tweets as “disputed” and “false.” As portions of the public embraced this new censorship, platforms moved from merely fact-checking presidential posts to deleting them entirely and threatening to ban other elected officials with whose policies they disagreed.

Eventually, even the news media lost its deferential treatment. Much like China’s state-controlled internet, mainstream American newspapers have begun to confront having their accounts suspended and posts deleted for reporting stories that the platforms deem “harmful” to society. Public figures in the US are even beginning to enjoy the UK’s model of super-injunctions as they ask social platforms to ban links to news coverage that they find embarrassing or politically harmful.

In short, every corner of American society today, from the presidency to the fourth estate, is now beholden to the acceptable speech rules and censorship powers of Silicon Valley online; yet as a society, we have little visibility into the rules that now govern the digital public square.

Addressing Social Media Censorship Is the Free-Speech Issue of the Decade

To some, Silicon Valley’s newfound emphasis on combating “misinformation” and arbitrating “harmful” speech might seem like a positive development. After all, threats of violence, racism, sexism, doxing, sedition, and harmful medical advice are damaging to society. Yet billionaires who can silence presidents, a government 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.

Why do the speech rules of social media companies matter? They matter because the technology to enforce real-time society-scale censorship has arrived without the corresponding societal processes and agreements over what should be censored. For over 200 years, Americans have argued over how to define “acceptable speech” and experimented with almost every form of censorship—to no avail. Social platforms are not merely implementing preexisting speech rules passed by Congress or voted on by the American public; they are deciding for themselves what should be permissible speech in the US, without any feedback from the nation’s citizenry.

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 that a handful of companies now decide the online speech of the entire planet. With a few lines of code, a person or an idea can simply vanish from the digital world, while AI algorithms are increasingly being turned loose to try to 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.

The ease with which we can now censor masks the simple fact that the most important question of all remains unanswered: What should be censored?

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 and create an era of "censorship without representation."

What Supreme Court Justice John Marshall Harlan called in 1971 the "intractable" problem of defining America's acceptable speech has eluded all consensus, so we have effectively given up as a nation and left it to private companies to sort out on their own. Silicon Valley has been entrusted to decide which beliefs and ideas are acceptable to American society and which it believes are "harmful" and to enforce those rules on our new digital public squares. 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.

The transition of the nation's public squares into the walled gardens of social media mean that our democratic debates are increasingly decided on the equivalent of private digital property where the concept of free speech does not apply. The First Amendment holds only that the government may not restrict speech; it does not apply to social media platforms, even when they restrict the speech of government itself. In turn, Section 230 of the Communications Decency Act empowers those platforms to censor or not censor at will and exempts them from nearly all legal liability for their actions, going so far as to strip the states of their right to narrow any of its provisions.

At the same time, Section 230 is not immutable. Notably, it was amended to narrow its protections governing sexual trafficking, meaning that there is a precedent for modifying its provisions.

What options does Congress have? Democrats largely want to increase social platforms' roles in governing American speech. Republicans largely want to eliminate their censorship powers and primarily emphasize the role of competition. Libertarians suggest creating new platforms. None of the three approaches is likely to change the status quo.

While they differ on their proposed solutions, all sides of the debate over social platform moderation agree that the current state is untenable. Yet without a greater understanding of how well the current system is working and its unintended consequences, it is impossible to know what the best approach might be.

What is missing is transparency into the inner workings of today's social platforms and data that would permit external evaluation of their impact on society by journalists, researchers, policymakers, and the public. A bipartisan first step, therefore, would be to require that in exchange for the unique Section 230 protections that they enjoy, social platforms should be required to provide a set of critical public data sets that would enable, for the first time, external scrutiny of their daily impact on democracy.

PART 1: THE STATUS QUO IS FLAWED

The challenges of social media moderation and its impact on society have attracted bipartisan interest. Three primary solutions have emerged: increase moderation to remove all objectionable material; repeal Section 230 to reduce moderation; or create alternative platforms that are less moderated. All three solutions have existential flaws.

Increasing Moderation Risks Abuse

Democrats have increasingly called for social platforms to do more to remove what they view as harmful material, from hate speech to misinformation.

At one hearing in October 2020, Sen. Ed Markey (D-MA) claimed that the true censorship problem was not that social media companies censor too much but that they don't censor enough: "The issue is not that the companies before us now are taking too many posts down. The issue is they are leaving too many dangerous posts up."

At the same hearing, Sen. Chris Coons (D-DE) demanded that social platforms further censor political "misinformation" such as "climate denialism." He warned: "I'd urge you to reconsider [your current censorship criteria] because helping to disseminate climate denialism, in my view, further accelerates one of the greatest existential threats to our world."

Sen. Richard Blumenthal (D-CT) called on social media to restrict "destructive" information. He said, "I recognize the steps—they're really baby steps—that you've taken so far. The destructive, incendiary information is still a scourge."

At the same time, the history of American broadcast regulation offers a stark warning of what happens when private companies are called upon to act as societal censors: barring politicians and topics that they view as harmful to their business interests, silencing antibusiness speech and whistleblowers as misinformation, and being forced to promote presidential policies and silence criticism of government. The long track record of abuse over the past century, when America's broadcasters were called upon to play a similar role, should give pause to Democrats' enthusiasm for Silicon Valley to play censor.

Efforts to Repeal Section 230 Could Worsen Censorship

Republicans have often responded to social media censorship by calling on Congress to end social media's Section 230 protections. In 2020, President Trump signed an executive order removing Section 230 protections from social media companies that censor political speech. Sen. Josh Hawley (R-MO) introduced legislation in Congress that would do something similar.

In 2018, Sen. Ted Cruz (R-TX) argued, "Right now, big tech enjoys an immunity from liability on the assumption they would be neutral and fair. If they're not going to be neutral and fair, if they're going to be biased, we should repeal the immunity from liability so they should be liable like the rest of us."

If the Section 230 safe harbor is eliminated, social media companies would be liable for the content on their platforms, similar to the way traditional media is regulated today. Yet this reform would likely result in more, not less, political censorship because social media companies would have to take immense steps to avoid liability, meaning far less approved content.

As Twitter CEO Jack Dorsey has explained, eliminating Section 230 could result in "increased removal of speech, the proliferation of frivolous lawsuits, and severe limitations on collective ability to address any harmful content and protect people online."

Rather than defend themselves from the inevitable wave of lawsuits if Section 230 were repealed, social platforms are more likely to enact draconian new content rules that eliminate much of the free-wheeling speech for which they are known today. Instead of freer speech, the result would be far more limited speech.

The immense costs associated with warding off liability would likely cement further the oligopoly status of the existing social media giants. Compliance costs, including building up the necessary technological and legal capacity, would likely be prohibitive for potential competitors looking to launch. The competitive moat generated by regulation costs is likely a major reason that social platforms support government regulation, both in the US and abroad.

In fact, after Europe implemented the costly new privacy regulations known as the General Data Protection Regulation in 2016, social media giants further consolidated their power and presence. Eliminating Section 230 would, therefore, likely backfire and result in more censorship and more social media oligopoly power.

Free-Market Solutions Are Insufficient

Free-market advocates have long argued that market competition is the best response to social media censorship. If you don't like censorship, they argue, create another platform with more freedom. This position has been part of the founding ethos of Silicon Valley.

In the words of the free-market Pacific Legal Foundation: "Do social media sites have a legal obligation to allow equal access to all viewpoints? Do they violate the First Amendment if they exclude controversial speakers from their platform? Should the government step in to take corrective action? The answer to all these questions is a resounding no." If they censor, writes the organization in a representative view of this argument, "we can vote with our wallets and our time" and support "alternative platforms like Parler. That's how a free market operates."

This argument is flawed when you consider how the tech industry, like the American motion picture, radio, and television industries before it, has adopted largely uniform acceptable speech guidelines that all companies, including new entrants, are effectively required to follow or risk being disconnected from the digital ecosystem. When one company bans a person or class of speech, its peers move in lockstep. As Twitter and Facebook banned Donald Trump, platforms 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.

The fact that Twitter and Facebook competitors Snapchat and foreign-owned TikTok 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 that everyone, even foreign companies, are required to follow. It also suggests that antitrust action would do little to reduce censorship if even foreign companies are now adopting the same joint set of content rules.

Companies that don't follow these rules are effectively removed from the internet.

Consider the story of Parler, which tried to follow this free-market advice to compete with existing platforms by creating an alternative that it saw as more pro–free speech, with light moderation and few banned topics. In the aftermath of the events of January 6, 2021, which Parler was viewed as helping to foment, Apple and Google responded by banning Parler from their app stores. Then Amazon pulled its web hosting, taking it off the internet completely.

Parler was able to return to the marketplace only after agreeing to implement new moderation rules, more closely aligned with those of the rest of the tech community. In short, what were once individual company policies governing acceptable speech have now become the rules of the internet that all companies must abide by in order to connect to today's Silicon Valley–controlled digital ecosystem.

PART 2: TRANSPARENCY NURTURES CREDIBILITY

Social media companies today routinely restrict posts and suspend, ban, or demonetize 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 change over time. Twitter originally claimed that it was banning sharing of the *New York Post*'s Hunter Biden story because it was "harmful." It then said that it was a violation of its hacked materials policy, before changing its story a third time to say that it violated its personal information policy. After a public outcry, it finally removed the ban and admitted that it was "wrong" and "a total mistake."

Twitter's ever-changing justifications for censoring the *Post*'s reporting remind us that as social platforms have grown from niche websites, their moderation policies have not evolved to befit their increasingly central roles in today's society as the gatekeepers of our news and information, the arbitrators of acceptable speech and ideas, and even the gateways to our elected officials. In essence, they've never moved beyond their startup roots of treating content moderation as an ad-hoc process in which humans and algorithms cursorily glance at posts and make split-second decisions, with little concern about mistakes. For a small startup trying to keep egregious posts out of the headlines, such an approach might be a reasonable trade-off. For trillion-dollar companies that increasingly act as gatekeepers to the news itself, this ad-hoc approach is insufficient, given the impact of those decisions.

Instead, as a condition of the unique Section 230 protections they enjoy, social platforms should be required to provide transparency around their content moderation activities, forcing them to adopt far more rigorous processes, while improving the credibility of platforms by shining greater light on their invisible hand in our national debates.

Transparency Encourages Accountability

Controversy over social media censorship often derives from a lack of clear rules defining precisely what is permitted, preventing open societal debate about the acceptability of those rules and leading to uneven enforcement. Precisely defining the moderation rules of today's social platforms would permit a more informed societal debate and make it easier for users to understand whether their speech complies. Policymakers and the public should pursue transparency as the most important first step toward addressing concerns over social media censorship.

For all the concern over social media's community guidelines, content moderation, fact-checking, and advertising policies, few actual data points are available to evaluate moderation practices. Perhaps the public would actually agree with most decisions if they were made in a transparent and objective manner. On the other hand, transparency may reveal that social platforms are egregiously censoring far more than generally realized. Either way, transparency is needed to make moderation decisions more accountable.

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. 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.

On paper, social media 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. 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? Transparency around the human and algorithmic moderation of today's platforms is urgently needed to answer these questions and empower democratic debate over how these platforms are shaping our societies.

Social media companies openly acknowledge the difficulty of their work. In 2017, Monika Bickert, Facebook's head of product policy, noted 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."

How would Americans react if they fully understood the disproportionate impact that censorship can have on underrepresented voices or knew the unevenness in how the platforms apply their rules? Would the public have supported Facebook's previous policies of allowing graphic threats of violence against women, gender-based attacks on women drivers, and race-based attacks on minority children and providing a special marketing category for "Jew haters" or allowing its recommendation algorithms to encourage anti-Semitism?

None of these insights was provided by the companies themselves; they were all leaked or discovered by researchers outside the companies, shining light on their practices. Yet social media's centrality in modern life means that we cannot depend on these chance revelations; companies must be compelled, based on the unique regulatory treatment that they enjoy, to provide sufficient transparency to enable public debate over their policies.

In order to accurately examine the impact of social platforms on society, we need data that capture the daily functioning of our modern public squares. Full transparency can turn seemingly political and arbitrary censorship decisions into objective and fair content moderation policies through a clear understanding of their rationale and basis and a democratic debate over their implementation.

Amending Section 230 to Usher in Basic Transparency

Social media platforms today have no legal obligation to provide even the most basic transparency around their moderation policies, how they enforce them, how they train their algorithms, 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, social media companies be required to provide 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. By lifting the veil on moderation decisions, policymakers and the public can understand and debate free-speech standards fairly applied to all.

Social media moderators should be required to clearly document in plain language the rationale behind each enforcement action. Such explanations should cite the specific policy violated, along with supporting evidence and precedent, and clearly state why the moderator believes the post to be a violation. 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 language, are enforced evenly for all users, and provide a detailed personalized explanation with each enforcement action (not a generic template).

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 the option of a live chat with a real human moderator to appeal the decision. Such an appeal process should have a guaranteed turnaround time of less than 12 hours and a bias toward restoring the post.

Transparency Can Reduce Capriciousness

Requiring moderators to clearly and thoroughly explain their decisions rather than simply clicking "keep" or "remove" would force them to carefully justify their verdicts according to policy and precedence rather than ad-hoc gut feeling. If social platforms must apply moderation decisions to a transparent public framework and publicly explain every decision, those verdicts and rationales can inform and be shaped by democratic societal debate.

A documentation trail that users can optionally share with external researchers to evaluate the consistency of social platforms' decisions and the degree of accuracy with which policies are being implemented will also reduce the urge to capriciously censor. If moderators must clearly explain each decision and cite appropriate policy and precedent, it encourages them to adhere more closely to the rules and spend more time considering precedent-setting cases. The platforms themselves will no longer be able to maintain separate unpublished policies and exceptions for politically sensitive or connected groups or quietly treat certain users differently without those trends being observable to researchers.

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

Ten Databases to Generate Social Media Censorship Transparency

Bringing transparency to social media censorship decisions begins with the data necessary to evaluate platforms' actions. Below are ten data sets that Congress can demand from social media companies to provide critical insights on censorship practices and reveal potential biases and problems on the path to more objective and clear censorship criteria.

1. Public Algorithmic Trending Data Sets

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, while Facebook was filled with the smiling faces of people dumping buckets of ice water over their heads. A public data set capturing how public posts (thus avoiding the privacy issues of private posts) are being prioritized or deemphasized by these algorithms across user communities and over time would provide insights into explicit and implicit biases in these algorithms and provide greater visibility into what the public is and is not seeing.

2. Automatic Database of Public Post Violations

Given that all tweets are publicly viewable and already accessible to researchers using Twitter's APIs (application programming interfaces), there would be few privacy implications in requiring Twitter to provide a public database of all tweets that the platform flags each day, along with a description of why Twitter believes that each tweet is a violation of its rules or disputed by a fact-checker. Such a database would permit at-scale analyses of the kinds of content that Twitter's moderation efforts focus on. At the same time, it would allow the public to compare violating tweets against the rest of Twitter, evaluating whether the platform's removal efforts are evenhanded and effective.

The Lumen DMCA (Digital Millennium Copyright Act) 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 social media posts, all removed content could be indexed into a similarly structured database.

Social companies would likely argue that this transparency requirement 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. Lumen's preexisting solution to redact full URLs shows that minor tweaks can avoid this pitfall.

Politiwoops already archives in a searchable database tweets by public officials that they've later deleted. While Twitter suspended the project's access in 2015, it eventually restored it and has allowed it to continue. However, the archive contains only those tweets that politicians themselves delete, not content that Twitter removes as a violation—though, in most cases, Politiwoops should catch such content since Twitter typically does not delete tweets but rather locks an account until users delete offending tweets themselves. However, in those cases, there is no explanation that the deletion was forced by Twitter rather than voluntarily removed by the user and there are no details about why Twitter felt that it was a violation.

One could imagine a system in which removed posts by public figures are archived in their entirety, similar to the Politiwoops model, while removed posts voluntarily submitted by ordinary citizens are archived akin to Lumen, allowing the public to see basic details, while journalists and researchers can access all submitted 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 for the deleted post. Some users might not, while others might be glad to share the selectors to help capture how policies are affecting various demographic groups.

3. Database of Deleted and Exempted Protest Posts

Protest marches are increasingly being organized over social media. As platforms extend their censorship of these posts, 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 in 2020, Facebook touted its removal of Covid-19 "reopening" protests that did not require social distancing, yet quietly waived those rules for the George Floyd protests. Having a centralized database of protest posts removed by platforms as well as those exempted from its rules would go a long way toward understanding how much the platforms are shaping the offline discourse.

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's post 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" or otherwise declines to take action. 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 that the companies rule are not violations would offer critical insights into how evenhanded they are and where their enforcement gaps lie.

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 that the platform ultimately determined were not a violation and allowed to remain.

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. Greater transparency around differing enforcement would push companies to codify these differences in writing, rather than quietly waive their rules.

4. 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 theoretically have 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 or potentially conflicted sources?

To create transparency around their fact-checking removals, platforms should be required to compile a database of every post they flag as being disputed by a fact-checker and make it available to journalists and researchers. For public posts such as those on Twitter, this would be trivial, but for platforms like Facebook, this would pose a privacy challenge. Beyond allowing users to voluntarily submit their own removed posts, another 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 the climate-change fact-checks published over the years, which of them 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 outsize 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. During the pandemic, public health officials could use these data to flag emerging contested narratives or remove outdated guidance by focusing on the most heavily used fact-checks.

5. Database of Private Post Violations by Journalists and Politicians

Most social platforms offer 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 nonpublic 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 those of other users, given their outsize 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 that they received regarding the violation. Having a single centralized database of such removals would make it easier to understand trends in the kinds of content that platforms are most heavily policing and whether there is public agreement with the platforms' decisions.

Having such a historical record would also allow researchers to look back at the impact of outdated guidelines. For example, Facebook's Covid-19 guidelines long prohibited many claims regarding vaccine-induced side effects, even after widespread medical community documentation of rare blood clotting. A historical removal database would allow researchers to look back to see if there was a steady stream of clot-related posts that Facebook had been deleting that could have served as an early warning to the medical community.

6. 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. While these attributes are imperfect, the fact that the companies make them available for ad targeting suggests that they believe that 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 immigrant religious affiliation? Do the companies’ enforcement actions appear to disproportionately affect vulnerable voices?

Companies should be required to provide daily or weekly summaries that list each specific moderation policy and fact-check 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 or fact-check. 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 the policy violations for that group.

This is critical in understanding whether policies are inadvertently disproportionately affecting certain groups, such as women or minorities. For example, are hate-speech policies inadvertently being enforced more often against women or minorities? Are fact-checks being enforced more against certain demographics?

7. Increased Access to Facebook’s Fact-Checking Database

Facebook provides an internal dashboard to fact-checking organizations that lists the posts that it believes may be false or misleading. 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 that 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.

8. Increased Access to Facebook’s Research Data Sets

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

9. Database of Posts Referred to the Legal System or Removed Because of Offline Harm

Many of the “community guidelines” enforced by social platforms are, at least on paper, also violations of US 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 that harassment and threats of violence should not be reported to officials if they believe that 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 US laws adhere to reality.

Companies routinely remove content such as 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.

10. Self-Submission Database of Private Posts

For private content, social platforms could be required to offer users a one-click button to voluntarily 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 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 nonconsensual 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 that the user is willing to share.

PART 3: WIKIPEDIA AS A TRANSPARENCY MODEL

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

In contrast, Wikipedia offers an example of how to transparently confront content decisions as its contributors and administrators publicly debate each day everything from what warrants inclusion to how it is presented and what evidence is cited. Thorny content debates play out in full public view on topics from the most mundane to the most controversial.

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

Wikipedia incorporates at least three editing strategies that social media platforms can adopt and adapt to make their moderation decisions more transparent and accountable:

- All content rules and precedents are public
- Archived public history of all content decisions
- A robust public "Talk" page that archives the entire decision-making process around key decisions

Clear Consensus-Driven Rules Can Bring a More Objective Standard to Social Media Moderation

Wikipedia editing is based on a combination of transparent public rules and public consensus-driven debate, which help minimize subjective editing and overcome the type of ad-hoc content decisions that plague social media moderation. According to Wikipedia:

> Wikipedia's policies and guidelines are developed by the community to describe best practices, clarify principles, resolve conflicts, and otherwise further our goal of creating a free, reliable encyclopedia. Although Wikipedia generally does not employ hard-and-fast rules, Wikipedia's policy and guideline pages describe its principles and agreed-upon best practices. Policies are standards all users should normally follow, and guidelines are generally meant to be best practices for following those standards in specific contexts.

Importantly, all of Wikipedia's policies, guidelines, and precedents are public. In contrast, social platforms have long resisted releasing their detailed internal moderation guidelines, publishing only vague rule lists that emphasize the importance of subjective interpretation by moderators. They also maintain myriad ever-changing unpublished exceptions to their rules that change in step with political sensitivities and administration changes in each country. The only glimpses of these more detailed guidelines have come in the form of leaks by employees. Some of these leaks have prompted public debate and changes to policies, including those permissive of racism, anti-Semitism, and violence against women, illustrating the critical importance of transparency.

Such an environment inevitably leads to confusion and concerns of bias when users have no way of knowing what speech is acceptable or banned, when what is acceptable today is prohibited without warning tomorrow, when one user is permitted to say something that another is barred from uttering, and when even permitted content can lead to permanent banishment without the possibility of appeal because of human or machine error. Adding to the confusion, companies typically refuse to comment when journalists ask whether a given statement is or would be a violation of their rules, offering that the only way to find out would be to post it and see whether the post or user is removed.

Social media users should know—before they post about climate change, election integrity, the property holdings and statements of public figures, or any other contested political issues—whether they will be censored, based on the violation of clear platform rules. The status quo, where social media companies censor after the fact, based on opaque ever-changing and unevenly enforced guidelines, breeds confusion and accusations of bias. Most important, the lack of visibility into platform rules deprives democratic societies of the opportunity to discuss and debate those rules, such as whether in 2017, Facebook's rules permitting glorification of violence toward women were acceptable.

Archive of Moderation Decisions Allows for External Review

Wikipedia archives all edits by date and contributor. This process allows users to see how the content of a given page has evolved over time. Users can easily see what content was added or removed, down to the punctuation, for a given article, along with an explanation as to why. Producing this record allows for external review of all changes. The community can notice if an edit appears to deviate from the site's policies or conflicts with other known information and can instantly revert those changes, while external researchers can flag concerning trends.

Editing activity is also archived by contributor, so other users and even external researchers can examine a user's sitewide edits to flag potentially malicious or rules-breaking users. Any given article edit can therefore be considered in the context of that user's entire history on Wikipedia to lend context and confer reputation and trust to those users whose edits are rarely reversed.

Social media companies could follow this model of publicly archiving moderation decisions for public posts (and private posts with the permission of the posters) to allow for greater analysis and understanding of their content policies in practice. Such a model can help users better understand the integrity—or concerns thereof—of moderation actions.

Public Record of Debates and Decisions Allows Users to Understand Policies in Practice

Examining Wikipedia's Talk pages illustrates how its decentralized community of users come together to discuss the different sides of a controversial argument before a final decision is made on whether and how to include it in an article. Users cite Wikipedia's rules and precedents and compile external evidence to argue for their desired outcomes, while the most contentious decisions may even be put to a vote. This permanent public record of the decision-making process and its supporting evidence ensures transparency for even the most contentious of debates.

Social media companies could similarly reduce controversy around their moderation decisions by making available moderators' arguments for and against removal or retainment, demonstrating the thinking and objective rationale behind the ultimate decisions. This would include a detailed explanation of the moderator's thinking, citations to supporting evidence and precedent, and any discussions with other moderators or supervisors regarding the decision. Such transparency would help to convince, rather than merely inform, users of moderation decisions. Just as judges in a court of law preface their legal opinions with an explanation of precedent and logic to offer justification, social media platforms can make public their rationale for moderation decisions in order to demonstrate that they were arrived at logically and evidentiarily, not arbitrarily.

Browsing these Talk pages, especially those governing controversial topics, offers a glimpse into just how adversarial content debates can become. Some debates are rancorous; but most are more collegial, in the vein of opposing lawyers in a courtroom. From minor disputes over a citation or reference, these debates can rage over far more deeper questions, such as whether the subject of a page is a "prominent and influential scientist with wide community support" or a "fringe pseudoscientist who claims to have conversed with aliens." 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 how our smart speakers and phones answer our questions.

The way in which Wikipedia presents 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, 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.

Only later were the allegations against Biden expanded and given their own page. The Talk page for the entry shows just how divisive and controversial the editors found mentioning the accusations at all; at one point, a group of editors voted to delete the entry, but since there was no consensus under Wikipedia rules, an administrator left the page intact. Similar robust discussions in the public domain by social media moderators, with a bias toward content inclusion, could help overcome many of the criticisms associated with social media censorship by opening them to public debate and discussion.

Wikipedia Still Has Bias, but at Least This Bias Is Transparent and Measurable

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 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," 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.

Unlike on social media platforms, all this debate and editing is recorded for posterity, allowing the public to see whether a given debate was contentious or unanimous.

As with many online platforms, the demographics of who contributes to Wikipedia have historically hardly been representative of society at large. The site's historically majority male editors have, over the years, led to a site that has minimized the role of women in STEM fields. Moreover, as efforts were launched to better represent women scientists on Wikipedia, some of those editors moved swiftly to delete entries or incorporate them into the entries of their husbands, arguing that many women scientists weren't noteworthy enough to warrant their own Wikipedia entries. Similar concerns have been raised about Wikipedia's representation and coverage of other underrepresented groups, such as racial minorities.

The backlash from some of Wikipedia's 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.

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, while MSNBC is, in Wikipedia's eyes, a reliable, neutral, and trustworthy source for all topics, including politics. 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." Once again, Wikipedia's transparency means that these decisions are available for public debate, whereas the internal lists used by social platforms are highly secretive.

All these factors, from the demographics of its contributors and administrators to its rules, processes, and norms, embed various kinds of bias into Wikipedia's pages. Yet Wikipedia's transparency means that all these biases are available for study, discussion, and the development of mitigation strategies, rather than hidden from view, as they are on social platforms.

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 argued 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 these trends and open them to public debate. Indeed, the public debate and rule changes that have followed past leaks of internal Facebook 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.

CONCLUSION

America's two-century struggle to balance the First Amendment with the desire to constrain "harmful" ideas reinforces the simple truth that in a society as diverse and independent as the US, anything short of unfettered speech yields only an endless and intractable divide.

Over its more than 200 years, the United States has experimented with almost every model of censorship. Early attempts focused on regulating speakers; 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 turned into global rules. Allowing states agency to define acceptable local speech failed to prevent conflicts, as states attempted to silence speech from afar, while centralizing power meant that 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 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 once again reverted to the privatized censorship model of early broadcasting—this time, empowering private companies with near-absolute censorship powers through the creation of Section 230.

Transparency alone cannot solve our diverse and divided nation's disagreements over the ideas, beliefs, knowledge, and speech that should guide our democratic debates over our shared future. What it can do is transform today's closed and seemingly capricious rulings into a public process—akin to our legal and electoral systems—that can be scrutinized and publicly debated. We still may not agree with the outcomes of social moderation, but we can observe and ultimately influence the process.

Most important, by transforming content moderation into a public and observable process, we create transparency that sheds light on inadvertent biases and allows the nation's citizenry to weigh in on the rules that increasingly govern our public spaces.

In the end, transparency is merely the first step toward a broader portfolio of changes, but by allowing us to observe for the first time the inner workings of social platforms, it empowers the public debate upon which nothing less than the very future of democracy depends.

To view the full-length research report behind this analysis, including citations, please visit:

https://assets.realclear.com/files/2021/10/1892_leetaru-social-media-digital-censorship-and-the-future-of-democracy-working-paper.pdf

RealClear Foundation