CATO
POLICY
ANALYSIS

May 21, 2024 Number 974

A Guide to Content Moderation for Policymakers

By David Inserra

EXECUTIVE SUMMARY

Content moderation represents the policies and practices that companies use to express their own preferences and to create the kind of online space that is best for their interests. Government policies that interfere with these content decisions not only harm the rights of private actors but also are likely to cause harmful unintended consequences and chill innovation. While prominent social media platforms may be biased and imperfect, the government cannot solve these problems and will only make them worse.

Policymakers worldwide are increasingly advancing policies related to content moderation. From the left, there are efforts to stop hate speech and misinformation, as seen in New York's Online Hate Speech Law and the European Union's Digital Services Act. From the right, there are efforts that try to force social media companies to host content from certain political speakers or viewpoints, as seen in legislation in Texas and Florida. Despite the intensity of these concerns—some of which may be valid—efforts to regulate content moderation often reflect a lack of understanding of how content moderation works.

Policymakers should understand that content policies are rules, protected by the First Amendment, which organizations use to create their preferred spaces. These policies must work when applied to billions of different pieces of content. No matter the principles a platform holds and no matter the wishes or intentions of policymakers, these companies need policies that they can implement effectively and consistently, something that government regulation generally undermines. Content moderation also comes in all shapes and sizes, including an increasing interest in giving users greater control over their experiences online. Government restrictions and requirements will likely prevent future innovations that better serve and empower users. Instead, those who value a culture of free expression should engage with current and emerging social media platforms to push back against the prevailing norms that are critical of expression and instead affirm the importance of giving people a voice.

DAVID INSERRA is a fellow for free expression and technology at the Cato Institute.

INTRODUCTION

Before contemplating the practices and policies of moderating content, it is important to first understand the scope and scale of online content.[1] Consider social media and video sharing platforms. Over two billion people were active on Facebook every day in 2023.[2] Users posted around 27 million new TikTok videos every day in 2023.[3] More than 500 hours of video are uploaded to YouTube every minute.[4] And this only continues to grow as social media use increases and the number of users expands.

Furthermore, beyond these most obvious examples of websites hosting content, news organizations and blogs often allow for readers to post comments to discuss current events.[5] Wikipedia is built by users who create and update entries. Businesses such as TripAdvisor, Yelp, and OpenTable rely on user reviews and comments to provide recommendations and resources to travelers and patrons. Amazon, Etsy, and countless other online retailers set standards for what products can be sold and how they can be marketed, as well as hosting comments and reviews of the sellers and their products. Online video game stores such as Steam allow users to sell their own independent games and to also craft modifications (mods) that can superficially or significantly change a base-model video game. While much of this paper will address larger social media companies, it is important to remember that many other businesses rely on user-generated content and therefore have content moderation policies.

CONTENT POLICIES

In response to the massive growth of content posted online, companies that host third-party content (i.e., content posted by users of an online service rather than content posted by the service itself) need to decide how and when to remove content that may be harmful, unlawful, offensive, or otherwise objectionable. Section 230 of the Communications Act, passed by the Communications Decency Act of 1996, clarified that websites that host third-party content are not responsible for such content and the act enables them to moderate this content without being legally liable for their moderation activities.[6]

Unlike traditional media, where each printed newspaper or each aired television program must be published by the media company, the authors of Section 230 recognized that the internet was different in that users, not internet companies, were the ones posting massive amounts of content online. And so rather than hold websites and internet service providers liable for the content posted by internet users, Section 230 simply asserts that it is the users posting content who are liable for their own speech.

A further important distinction is that, while Section 230 was instrumental in allowing user-generated content and the websites that host it to flourish, it does not replace or override the First Amendment and property rights of website owners to exercise their editorial discretion over what they will not allow users to post on their services, in the same way that a website developer may refuse to advance a message with which they disagree.[7] And so companies have crafted rules, community standards, guidelines, and other forms of content policy to create their desired online space.[8]

Box 1

Glossary of Terms

Content: posts, comments, hashtags, pictures, videos, audio recordings, files, reviews, advertisements, product listings, and various other forms of online expression.

Operating or moderating at scale: the challenge of working with a massive amount of content.

Community standards, community guidelines, community policies, or rules: established policies for what is allowed on a given platform and what the punishments are for violating those policies.

Actioning content: a generic term for enforcing against a piece of content or user.

What Content Policies Are

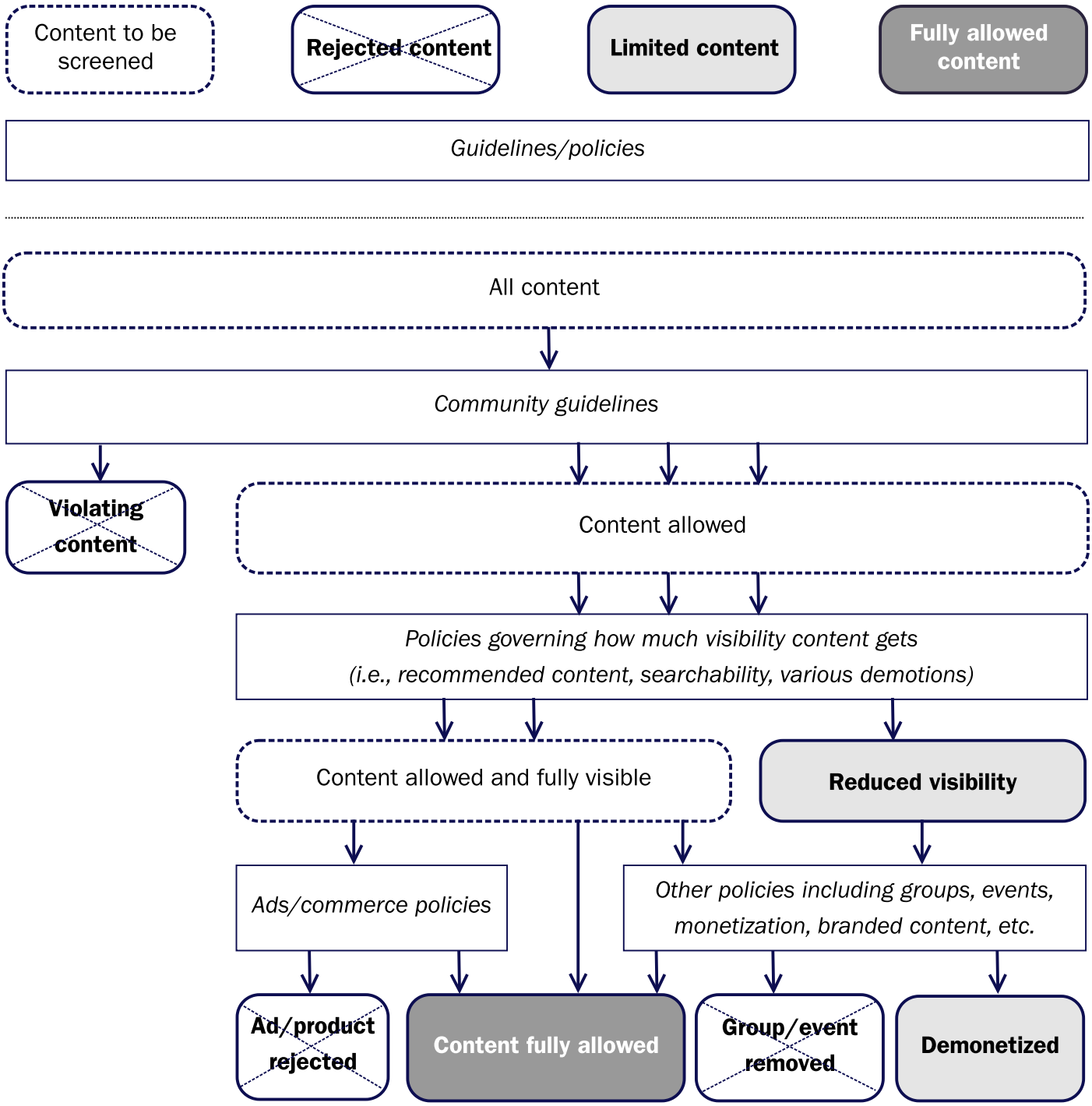
Content policies are the established rules for what can and cannot be posted on a given platform. They are the rules that companies employ to decide what content users are not allowed to post on their websites, and these rules establish when content is to be taken down or how users who violate these rules will be penalized. While a lot of attention is paid to individual content decisions, those decisions are usually the result of enforcing a set of rules that were previously decided and documented in some public form; however,

many platforms are not entirely transparent about the specifics of their rules.

Platforms often have several tiers of content policies: a least restrictive core policy that applies to all content; another set of policies governing recommendations, feeds, or timeline decisions; an even more restrictive policy for what kind of content can be contained in an ad or what kinds of goods or service can be sold online; and other restrictive policies governing monetization, branded content, events, groups, or other topics depending on the platform.[9] Figure 1 shows generally how content may be subject to multiple tiers and types of policies.

Figure 1

The generalized policy layers of social media content moderation



Source: Author's experience and analysis of rules published by major social media companies.

Content Policies Have Grown

Originally, content policy rules on social media platforms were often short and basic, such as Facebook's original hate speech standard prohibiting "hateful, or racially, ethnically or otherwise objectionable" content.[10] Today, these standards have grown by leaps and bounds, especially for some of the larger companies. This increase is for multiple reasons: to create new rules to prohibit or allow specific types of speech, to craft unique rules to serve specific communities or groups, and to better flesh out existing rules in order to allow for more consistent and transparent content moderation at scale.

Indeed, companies cannot always predict how people will use their platforms, especially during particular political or social events, such as the COVID-19 pandemic. Thus, their policies evolve to address changes in the way users are using their platforms in light of changes in society or various situations that arise.

To this end, Facebook's content rules in its terms of service have grown from 292 words in 2005 to its current externally facing community standards of 18,662 words.[11] Twitter's 2009 rules were one page long—but now rules from X, formerly known as Twitter, have 17 different policy areas, most of which are as long as, or even much longer than, the original rules. One of the newer social platforms, TikTok, has 30 different policy areas in its community guidelines. It is also worth noting that these are merely the publicly available policy lines, some of which are high-level and broad. These often require more granular, confidential policies that guide how the high-level policies are applied at scale.[12]

As these policies have grown, so too have the ways that that content might be "actioned" or have some enforcement action be taken. These enforcement actions, penalties, or limitations include

- reporting the content to law enforcement;
- deleting the violating user's account;
- blacklisting or otherwise making content not postable on a platform;
- removal or deletion of content;
- demoting content so that it appears lower in users' feeds and is less likely to be seen;
- making the content not recommendable or not searchable;
- labeling the content as somehow false or wrong;
- adding outside information or context that users can see while reading a post;
- using interstitials—tools that blur the content or put it behind a warning screen unless the viewer chooses to see the content by clicking a button;
- restricting content to users that are over a certain age; and
- rejecting an ad or item for sale.

Box 2

Tide Pod Challenge

An example of the way platforms did not expect a certain user behavior and had to adapt their policies is the Tide Pod Challenge. While positive challenges such as the ALS [amyotrophic lateral sclerosis] Ice Bucket Challenge went viral on social media, so too have various harmful challenges, such as the Tide Pod Challenge, which dares users to consume a Tide Pod that is full of detergent and laundry chemicals. Platforms certainly didn't have "don't eat laundry pods" in their content policies, and so they had to adjust or add new policies around high-risk viral challenges.

Another example concerned COVID-19. Platforms did not have policies on how to manage government lockdowns, social distancing, users looking to get COVID-19 so as to gain natural immunity, users discussing various treatments for COVID-19, users dismissing the virality or severity of COVID-19, or users broadly or narrowly challenging the efficacy or safety of vaccines. Now, while many of these policies came under fire (such as enforcing government lockdown rules or harming legitimate discussions of COVID-19's danger relative to the costs of government restrictions), platforms certainly needed to address a variety of new types of content with varying degrees of harmfulness.

Why Content Policies Have Grown

Platforms typically encounter a wide range of potentially objectionable content—varying from nudity to regulated goods to hate speech. Each policy line must spell out exactly what content should be allowed and what content should

be removed or otherwise moderated. The explanations must be interpreted in a uniform way to ensure consistent application of the policy. And each policy draws a line that reflects the kind of content that each platform wants to allow or forbid its users from seeing or posting. Should hardcore porn be allowed? What about uncovered female nipples? Does it depend on the context or user? What about promoting the recreational use of drugs? Should anyone be able to see this? Does the type of drug matter? Is it hateful to say that one group is more intelligent than another? What about saying that one group is more educated than another? What about saying that the way one group is acting is dumber than another?

In addition to the variety of content issues presented to popular general-use platforms, there are also a range of context-related questions that a platform must consider. How do these companies handle content that is humorous or satirical? How do reviewers differentiate between humor and satire? What about content that condemns or is trying to raise awareness of some awful event? How clear does the condemnation need to be? Are some categories of content so harmful that they always need to be removed, regardless of the context? What often seems like a simple decision is covered by a series of intersecting policy lines that reflect a complex view of what content should be on any given service.

Given the need to operate clearly at scale—especially with millions of users posting content each day—some policies may be adopted because they represent the easiest way to consistently moderate content. This may result in curtailing speech, even though there are principled reasons to allow it. Some examples are illustrative here:

- **Violent speech in a nonviolent context:** Violent speech can be difficult to moderate because people use such speech in a variety of nonviolent contexts. "I might kill my ex." This phrase could reflect a person seriously considering murder. But it could also represent a hyperbolic way of dealing with a breakup. It could also be someone singing the hit SZA song "Kill Bill."[13] So how does a platform moderate this simple phrase? Do they always remove such content? Do they always allow violent speech if it is used in a humorous or musical context, even if such lyrics may be meant literally by some users? What are humorous or musical contexts? The principle of protecting artistic or humorous speech may be a good one, but platforms may choose enforceability over principle.
- **Viewpoint neutrality:** Services are under no obligation to be viewpoint neutral, and some may explicitly cater to certain types of users and viewpoints. But even if companies wanted to be viewpoint neutral, how do they define a viewpoint in a way that can be consistently applied? And can they do so without invalidating the rest of their content policies? Prostitution, pornography, and even child pornography are viewpoints. ISIS and the KKK each have a viewpoint. Pro-anorexia and suicidal content represent viewpoints. Allowing such content is certainly not what advocates of viewpoint neutrality intend, but it may be the result of applying such a principle to an online platform. And if platforms

Box 3

Meta's Known Questions

Meta's community standards contains the top-level policy line, but there are also expansive "known questions" that often instruct moderators on how to apply a given policy line. Users can get glimpses into these known questions in several decisions by Meta's Oversight Board, such as when, in 2022, the board took issue with the way the known questions defined "praise" for the purpose of praising terrorist, hate, or criminal groups. The board found that one of the definitions for praise—content that "makes people think more positively about" a designated group—was overbroad and not aligned with the community standards. These internal policies can have a significant impact on how the top-level policies are implemented by defining key terms, setting the boundaries of what is or isn't covered by a given policy, or by providing detailed examples. The lack of transparency around these internal rules, however, poses a challenge to users and researchers who are trying to understand how policies are being applied in detail.

Source: "Mention of the Taliban in News Reporting," Meta Oversight Board, September 15, 2022.

can't meaningfully define what is and what is not a viewpoint, they may simply disallow all speech that even vaguely could be considered a viewpoint, leaving social media banal and superficial for users in a state that adopts such a standard.

Who Determines Content Policies?

The values, principles, and enforcement considerations that go into content policies reflect the considered judgements of the companies as to what kind of speech they want to have on their platform and what kind of speech they want to prohibit. Furthermore, companies must craft policies they are able to consistently apply to thousands or millions of pieces of content a day.

At Meta, Twitter, YouTube, and other large tech companies, these decisions are often made through a fairly substantial development process, which attempts to make an informed decision regarding whether to allow certain types of content.[14] The process is more robust than, but similar to, a newspaper editorial board debating its op-ed policies or a bookseller deciding which books to stock. At other platforms, their content policy development process may be less detailed or reflect a more decentralized approach to rulemaking that gives greater power to communities or users. But regardless of the specific deliberative process, all content policies fundamentally exist to create the type of online space that any given company wants to provide to users and that will attract advertisers.

It is important to note here, however, that just as newspaper editorial boards may be flawed or biased toward a certain party or viewpoint, the policy development process is also imperfect and open to bias. This bias can be explicit, like an editorial board simply wanting to advance certain narratives and speech, or it can be an unintentional, unconscious, or structural bias. While explicit bias and preferences are obvious—for example, Meta does not allow pornography while X does—large platforms' centralized content policies and the massive amount of expression they host creates an incentive for parties to lobby the platforms to change their content policies.

As with public policy, interest groups have pressured platforms to set their content policies in ways that align with the groups' views. Lobbyists for sugar growers can organize and succeed in acquiring a sugar subsidy because the benefits they want to acquire are concentrated and significant for them, but the costs to the average consumer are dispersed, creating little incentive for consumers to organize against the subsidy. In the same way, various interest groups that really don't like speech they deem to be hateful, such as the Anti-Defamation League's focus on

Box 4

Meta's Community Standards Policy Development Example

At Meta, I was on the team responsible for developing the community standards from 2019 to 2023. One of the policy developments I ran responded to a recommendation from Meta's Oversight Board to allow speech that spoke positively about nonmedical drugs such as ayahuasca or peyote in traditional or religious contexts. This policy development surveyed the medical literature on the harms of various drugs that are used in religious contexts—known as entheogens—and spoke with various experts and groups ranging from drug legalization advocates, health professionals, drug regulators, traditional healers, sociologists and anthropologists focused on indigenous uses of entheogens, free-expression groups, and other related speakers from around the world. Meta's teams investigated how this type of content appeared on the platform and engaged in some polling of users in various countries. We also spoke with various internal Meta teams, including communications, public policy, legal, safety, and human and civil rights teams. With all this information and feedback, my team was able to create various options for how the policy could change to practically accommodate some or all of this type of content, with each option presenting benefits and challenges. We presented our research, options, and recommendation to our policy leadership in what we called a policy forum, where we ultimately adopted our recommendation to allow the promotion (but not the buying, selling, or trading) of a select list of entheogens.

Source: "Policy Forum Minutes," June 28, 2022, Meta Transparency Center, January 31, 2024.

what they consider antisemitism or Media Matter's efforts to combat what they deem "conservative misinformation," will endlessly lobby for more content to be removed by platforms.[15] Each new policy against hateful speech advances the goals, support, and funding for these activists and researchers, but the cost of reduced expression is felt across the platform in often dispersed ways.

The same could be said for the industry that has grown up around misinformation and disinformation. There are incentives for censorial actors in the way of funding, approval within the broader content-moderation community, and power over how content is moderated, but not for the skeptics and advocates of free speech. For example, no free expression activist groups have joined Meta's fact-checking program because it suppresses purportedly false information. Thus the program has an inherent selection bias—only organizations that believe suppressing "false" speech is a wise course of action sign up to be fact-checkers. In other words, only those willing to suppress speech have power over expression.[16]

In general, free expression groups have not been nearly as active and engaged in the crafting of content policy as supporters of speech restrictions. Free speech groups wisely defend the expressive rights of social media companies. But groups and experts that are hostile to expression dominate the substantive content policy discussion in a way that is similar to academia in America today.[17] Some of the most politically biased and expression-critical fields of academic research,[18] such as sociology, communications, anthropology, and similar fields, are the bread and butter of content policy development.[19] These academics and aligned interest groups are actively, consistently, and aggressively telling social media companies about the many harms of freer expression, effectively setting the norms of what speech should and should not be allowed. Other groups, generally on the political right, have completely disengaged with many tech companies, leaving social media firms with even fewer external viewpoints. The result is that the content moderation and policy field, also known as "trust and safety," is largely captured by viewpoints that are not friendly to expression in the same way that a regulator may be captured by relevant interest groups.

Additionally, social media companies face a great deal of pressure from governments, generally to remove more speech. Internationally, this is often via formal censorship laws and authorities ranging from broad content regulation efforts such as the EU's Digital Services Act to laws such as Germany's Network Enforcement Act (NetzDG) that specifically impose liability on social media companies for hosting what German authorities deem illegal speech—an approach copied by multiple authoritarian nations to justify the removal of dissenting opinions.[20] In the United States, open censorship is forbidden and so efforts at suppressing speech have largely taken the form of government censorship by proxy, where government actors have pressured, coerced, or funded private efforts to have companies suppress disfavored but lawful speech, most notably in the current Supreme Court case of *Murthy v. Missouri*, which is looking at the issue of whether the government coerced or too aggressively pressured social media companies into removing election and COVID-19 related content.[21]

"In general, free expression groups have not been nearly as active and engaged in the crafting of content policy as supporters of speech restrictions."

This external pressure to remove more speech exacerbates the problem of having trust and safety teams that—even according to tech executives—have ideological biases.[22] Looking at donations to political campaigns as a proxy for ideological viewpoint, we can see a strong preference for left-leaning viewpoints among many of the major tech company workers, and personal experience confirms the lack of ideological diversity in most trust and safety teams.[23] Thus, even among the many well-meaning trust and safety professionals, certain arguments, assumptions, and viewpoints are widely accepted, while others are not understood or even dismissed. Even putting ideology aside, the trust and safety teams are, as their name suggests, generally structurally focused on safety rather than expression. Their job is to remove harmful content and keep users safe. Crafting new policies to remove more speech is also a boon for career advancement, as such policies

show greater "impact" and garner wider peer support. These implicit ideological and structural biases impact the development of content policies in the way questions and challenges are framed, the type of research and outreach that is done, and the internal support or opposition for certain policy changes.

"Content policies are the established rules for what is allowed on a given platform and what the punishments are for violating those rules. Crafting such rules presents challenges, given the variety of content and contexts and the sheer scale of the content, thus leaving opportunities for bias and flaws."

Some platforms, however, mitigate this problem by exerting less centralized control over content policies and grant greater control to individual communities or users. These less centralized platforms offer some core policy rules governing the entire platform, while individual communities or groups may layer on additional rules.[24] For services such as Mastodon, Nostr, or Bluesky (services similar to X/Twitter), even greater control can be given to users depending on what tools and decisions their specific platform or servers provides.[25] These services are considered "decentralized" because they share common protocols to allow communication between different platforms and servers, but each maintains its own features. Importantly, users can move to any other decentralized server or platform while still maintaining their network, just as one can move from Gmail to Outlook but maintain the ability to communicate with others regardless of which service anyone is using.[26]

In sum, content policies are the established rules for what is allowed on a given platform and what the punishments are for violating those rules. Crafting such rules presents challenges, given the variety of content and contexts and the sheer scale of the content, thus leaving opportunities for bias and flaws. Such policies may attempt to rely on or balance certain principles, but they also must be applicable and enforceable at scale. Thus, the challenges of accurately enforcing content policies are where I will turn next.

CONTENT ENFORCEMENT

Content moderation is the systematic method of decisionmaking, whether by machines or humans, that determines if the content that a user posts violates the preestablished policies of an online service and directs how the service takes appropriate action on that content. In essence, it is the enforcement of rules to create the space that any given organization wants to create.

Enforcement at Scale

As mentioned earlier, the sheer magnitude of content that needs to be reviewed presents an insurmountable challenge for platforms. YouTube removed nearly 20 million videos in 2022. Twitter removed 6.5 million pieces of content in the first half of 2022.[27] Facebook removed more than 115 million pieces of content in the second quarter of 2023, not including 676 million fake accounts and 1.1 billion pieces of spam.[28] In most of these cases, the majority of people would agree that removal was the right call, such as clear praise of ISIS, calls to violence against a group of people, or doxing someone. If Facebook got its moderation right 99.9 percent of the time, that would still be nearly half a million pieces of non-spam content incorrectly actioned in 2023. If YouTube got 1 percent of its content removal decisions wrong, that's 200,000 videos wrongly removed in a year. Any enormous number multiplied by even a very small number still results in many errors.

Who Does the Enforcement?

Online websites and platforms can use different content enforcement strategies to determine who or what is actually doing the content moderation. When most people think of content moderators, they think of humans reviewing reported violations all day. In the early days of content moderation, or perhaps for smaller blogs or comment sections today, such reviewers may be the content

moderators.[29] TikTok claimed in 2023 to have "tens of thousands of moderators around the world" while Meta employed around 15,000 content reviewers in 2022.[30] But with vast amounts of content being posted every day, it is incredibly difficult and expensive for human moderators to review every piece of reported content, not to mention all of the unreported content that could potentially violate platform guidelines.

Thus, the other major form of content enforcement takes the form of various artificial intelligence (AI) technologies. Before the pandemic, many social media companies were increasing the use of proactive technologies to detect and delete violating content before users saw the content to minimize their perceived harms and deliver a better user and advertiser experience. Proactive AI technologies review large swaths of content that users post and assess how likely a piece of content is to violate the content policies. The pandemic cemented this trend, as many social media platforms simply had a tiny fraction of their human reviewers available in 2020. So, they turned to technology as the only way to review content. As a result, for some categories of content, 99 percent of removals were proactively detected and removed by technology (i.e., no human was involved, it was purely assessed and actioned by a machine). TikTok claimed a 96.5 percent proactive removal rate in the second quarter of 2023.[31] What this means is that rather than just reactively reviewing reported content, machines are responsible for detecting and removing nearly all the violations that occur on the platform. If 34 million TikTok videos are posted every day and nearly all are subject to review and removal by technology, then there are more than 12 billion opportunities for technology to reach the wrong answer every year.

As a result, human content moderators often work in coordination with technological solutions. For example, when technology assesses content as likely violating a content policy, it is not always used to delete content outright but may instead be used to route that content to reviewers. This hybrid approach benefits from the raw processing power of technologies as well as the context and understanding of language that only human reviewers have. It also helps reviewers by actioning the most graphic, most disturbing, and most clearly violating content without needing a human to look at it. On the flip side, reviewers are needed to help train the technology. The thousands of pieces of content that are actioned by reviewers every day for a given policy line are the fodder that the technology needs to better distinguish between violating and nonviolating content.

Box 5

Machine Learning for Content Moderation

Proactive content enforcement technologies work by being shown large amounts of content and being told what content violates standards or is allowed. Given enough examples, the technology learns what kinds of words, phrases, and imagery are associated with a violation. When given a new piece of content, the system uses its prior learning to determine how likely it is that this piece of content is violating. As with most things, we don't demand 100 percent certainty before acting, and so these technologies can be set to action content at different levels of certainty. But even at 99 percent certainty, when applied to millions of pieces of content, technology is going to get a massive number of decisions wrong.

Source: "How Technology Detects Violations," Meta Transparency Center, last updated October 18, 2023.

Decentralized Content Policy Enforcement Offers Users Choice

Alternatively, some services use diffuse content enforcement techniques. Reddit relies on subreddits to create communities with a specific culture and a set of additional rules. These rules are not enforced by a large Reddit-wide team of reviewers or technologies, but mostly by the volunteer moderators, or "mods," of that specific subreddit.[32] While there is some moderation and review by the central Reddit administrators, a great deal of the work is done by the community mods as well as "AutoModerator" tools that support mods' efforts to moderate a subreddit.[33]

For decentralized networks, like the recently launched Bluesky, users have even greater control over their experience. Bluesky gives users more power to self-moderate their conversations as well as the ability to opt in to various

moderation services.[34] These services label content as being violent, intolerant, spam, etc., and then users have control over how different labels by various labeling services are handled. For example, a user can choose to be advised about all content labeled as intolerance by a given labeling service, to reduce the prominence of and blur any content labeled as gore by another service, and outright hide any content labeled as spam by any service. Some content will always be taken down, such as child sexual assault material (CSAM). But outside a limited set of mandatory takedowns, the moderation or labeling services that a user chooses are highly customizable and can be changed at any time.[35] Such a system essentially enforces each user's preferred content policy.

Appeals

Many social media companies also include some sort of formal appeals process. These take different forms depending on the platform. Similar to standard review processes, these appeals may make use of human moderation, technical moderation, or some combination of the two.[36] Appeals, however, still consume reviewer resources, and so the same challenges regarding the sheer amount of content are also a challenge. During periods of high appeals or limited resources (such as what occurred during the COVID-19 pandemic), social media companies may not have the capacity to review appeals. Meta, for example, has also created an external oversight board that users can appeal to after failing an appeal with Meta. This oversight board has the authority to make binding enforcement decisions on any piece of content. That said, the board issued 75 decisions between the start of 2021 and January 2024, indicating that it has a limited capacity for appeals.

MISTAKES OR POLICY DIFFERENCES?

When users take issue with the moderation of some piece of content (such as a content removal or a failure to remove content) there are three possible reasons for what has occurred:

- **Innocent mistake:** A social media moderator or technology made an innocent mistake to remove content that should have been left up or else left up content that should have been taken down, per policy.
- **Biased decision:** A social media moderator or technology made a decision based on purposeful or indirect bias, or even external coercion.
- **Correct decision:** A social media moderator or technology accurately applied the policy, but users disagree with the outcome or policy.

As described above, the vast amount of content being posted and reviewed means there will always be a great deal of these innocent mistakes. However, it is not uncommon to hear politicians and regular users alike complain about specific examples of biased enforcement.[37] Of course, this is possible, as even with the many controls and layers of review that large social media companies have in place, human bias can sneak in. As the *Murthy v. Missouri* Supreme Court case has shown, government pressure can also inappropriately influence content moderation decisions. Furthermore, as moderation technologies are trained based on human decisions, it is possible that bias could seep into technological moderation. However, it is often the case that users simply disagree with the policy line. Several examples are helpful here:

- A user may post a manifesto of a school shooter on Reddit, trying to highlight the killer's evil mentality or understand the killer's thought process. However, Reddit and many other platforms have decided that they will not allow manifestos for various reasons, including efforts to combat copycat killers.[38] The poster may claim that the decision to remove their post was to silence their perspective or an inconvenient narrative, but the content was enforced against accurately per policy—a correct decision. The poster would be better off objecting to the policies that may be inappropriately limiting civic discourse or unduly punishing users.
- Under Meta's hate speech policy, users are prohibited from offering services that aim to change people's sexual orientation or gender identity. When a religiously traditional church, synagogue, mosque, or other place of worship posts about a support group or opportunities to meet with a religious leader for individuals who want to align their sexuality with their religious views, this content may get

taken down. The houses of worship may object to what they view as mistaken or biased moderators, but the content does violate Meta's policy. Rather than objecting to ad hoc mistakes or bias in the enforcement of Meta's rules, users would be better off taking issues with Meta's policy biases that may silence certain viewpoints and do so without complete clarity or transparency.[39]

While anecdotes can provide important examples of how content moderation isn't working perfectly, it is important to separate out why it's happening. A growing amount of conflict regarding content moderation is likely due to fundamental disagreements over what is and what is not acceptable online and the expanding size of content policies.[40] Furthermore, the lack of transparency and clarity by many social media platforms can make it difficult for users to clearly understand what content constitutes a violation. At a basic transparency level, the major platforms have internal policy guidance and designations that attempt to define key terms, such as dehumanizing speech or promoting hate. Users may believe that they are not violating the policy, but given the way that the internal policy works, their content may actually violate. Further complicating matters is that even if greater transparency were provided, drawing clear lines for enforcement at scale is often difficult, and for certain topics such as hate speech, definitions are inherently vague and open to disagreement.[41]

THE FUTURE OF CONTENT MODERATION

Increasing pressures are being placed on social media companies to create policies and moderate speech in competing and often mutually exclusive ways. For example, research by organizations such as the Future of Free Speech have found that most of the prominent social media companies have suppressed increasing amounts and types of "hate speech" on their platforms.[42] In principle, each company can define "hate speech" itself, as there is no uniform definition, but in a growing set of jurisdictions, governments are increasingly requiring and defining hate speech and other "harmful" speech standards for social media companies. Notably, this is the case in the German NetzDG, the EU's Digital Services Act, and New York's Online Hate Speech Law.[43] Whether due to internal preferences to limit expression, pressure from interest groups and civil society, or direct and indirect regulation by governments, most prominent tech companies have decided to progressively adopt policies that increasingly exclude mainstream speech and viewpoints, including hate speech policies that silence various traditional views of sexuality and gender; misinformation and fact-checking regimes that suppressed scientific discussions around COVID-19 and reporting in opposition to the Australian Indigenous Voice referendum; and various civic and social harm policies that prohibited protests against US lockdown policies or against Canadian vaccine mandates.[44]

"A growing amount of conflict regarding content moderation is likely due to fundamental disagreements over what is and what is not acceptable online and the expanding size of content policies."

One of the results of social media companies increasingly limiting the acceptable range of speech is the growth of new platforms that are creating new spaces. Rumble has grown as an alternative to YouTube; Twitter spawned alternatives such as Gettr and, after the purchase by Elon Musk, services such as Threads. Other decentralized or federated platforms, such as Bluesky and Mastodon, have grown as well. Social media users may be sorting away from the major social media networks and selecting alternative platforms, often moving toward an increasing list of smaller platforms.

(Elusive) Network Effects

Some argue that a force called "network effects" inhibits this potential sorting or fragmentation of users into smaller platforms. Network effects is the idea that large social media networks have strong staying power and are able to resist competitors because the larger the network of people using a service, the more valuable the service is to its users.[45]

Indeed, some even argue that there is not true competition within social media platforms because social media depends on users being connected to a network. As a result, existing platforms will always have a competitive advantage.

Nevertheless, network effects have not stopped the rise of new social media platforms that provide a desirable product, such as TikTok.[46] There is also evidence that during significant (real or perceived) changes to a large platform's content moderation policies, users do migrate to alternative platforms. This can be seen in the case of Parler, which according to research of user growth by the Stanford Internet Observatory, showed jumps in user growth "in response to political events in the United States and the choice by other platforms to label or remove content from prominent individuals, including President Trump."[47] Similarly, the purchase of Twitter by Elon Musk and the resulting changes in the platform's moderation policy have led some users to leave in favor of other platforms.[48] Thus, decisions by social media platforms to moderate more, less, or just differently can open, and already are creating, room for competitors that can provide a compelling alternative.

Alternatives to Centralized Social Media Content Moderation

Rather than social media users splintering off onto their preferred social media platforms, another potential future might be for existing or new platforms to devolve more of their content policies and moderation to users.[49] This could take multiple forms.

One form is the less centralized Reddit model, with some centralized rules and additional rules created and enforced by the subreddit communities. Another less centralized option would be for platforms to keep a central set of policies that are always enforced against, but to also give users increased control over their own experience. Users would be able to set their preferences for allowing or not allowing certain types of content and for determining what is prioritized in their newsfeed. Platforms could allow civil society organizations to even create "preset" moderation and feed settings that subscribers could choose to use. In this same vein, middleware—add-on applications that users could use in tandem with social media platforms that modify how content is served to any given user—could give subscribers and civil society organizations choice and control over their social media experience.

Yet another form of devolution would be the potential embrace of decentralized social media models, such as Mastodon, Bluesky, or Nostr, which allow communication between different social media servers but give users in each server control over their moderation and user experience. As noted earlier, decentralized services provide different tools that allow users to customize what kind of content they want to see. Similar to civil society moderation presets or middleware, the content moderation and labeling techniques that may become available could empower any organization to create tools or services that could add to the user experience online. Moreover, decentralized services allow users to switch between different services at any time while maintaining their network. Whatever form it may take, a decentralized, more user-driven future may await social media users if the market is allowed to function.

"Whatever form it may take, a decentralized, more user-driven future may await social media users if the market is allowed to function."

Indeed, there are multiple advantages of devolving powers to users. For the businesses operating such platforms, they will maintain a larger set of users and likely improve the experience and engagement of users. Rather than a one-size-fits-all approach, giving users or communities greater control over their social media experiences means that fewer users need to be punished as severely. Content that might be removed under current approaches might be allowed but will only be seen in certain communities, servers, or by individuals who proactively choose to see such content. Such an approach would also allow users and communities that want even stricter policies to apply those settings for their own well-being. Giving users greater control not only benefits businesses, but obviously benefits all sorts of users who might not feel like one set of content policies or algorithmic feed decisions are good for them. This could also create additional advertising opportunities for platforms.

Currently, some advertisers might not want to advertise—or may be pressured into not advertising—on whole platforms because a tiny amount of their ads may be placed next to potentially offensive or controversial content. With users given greater flexibility and control over their feeds, advertisers could similarly be given greater control over where to place their ads. Some might choose to only place their ads next to the least controversial and benign content, likely at a premium, while other advertisers would be willing to save money and have their ads appear next to content that some may find controversial.

"Rather than expecting intermediaries to protect us from every real or perceived harm online, we need more empowered individuals."

Of course, these platforms have the right to create whatever kind of space they want with their policies. But just as technologies continue to improve in response to evolving consumer demands, social media users and advertisers are demanding a better experience.

A devolved approach may also benefit society for several reasons. Online platforms are often accused of creating "echo chambers" or "information bubbles."[50] Such claims, however, do not show that online algorithms caused polarization rather than just supplying preexisting preferences. The research instead suggests that the algorithms that platforms use to recommend content to its users may offer up extreme views, but that such recommendations have a limited impact on the actual political views and behavior of users. So, for example, just because YouTube may recommend videos that are highly critical or supportive of gun control after users watch a video about firearms, those videos may affirm users' existing beliefs on gun control rather than cause them to hold more radical beliefs.[51] This is relevant for society because it means online platforms aren't one-way trips to radicalization and polarization, but rather expose individuals to a variety of perspectives. When individuals are punished or removed for posting violating content, their underlying beliefs do not go away. Instead, deplatformed individuals generally double down. This is made only worse by removing them from broad networks of relatively diverse perspectives and leaving them with smaller, more private, and ideologically insular platforms.[52] Removing increasing numbers of users for a growing set of content policies is likely doing more than ever before to create the very echo chambers and polarization that so many have worried about.

A more devolved and user-first approach to content moderation and recommendations gives users the ability to select their own experience but also remain part of a larger network where they hear from opposing viewpoints and can easily change their preferences or the communities they use. Trusting and empowering users will support a greater culture of free expression online instead of affirming the paternalistic elite panic that emerges every time a new technology expands access to expression, or the view that free expression is too dangerous for average people to exercise.[53] Rather than expecting intermediaries to protect us from every real or perceived harm, we need more empowered individuals, or what Greg Lukianoff and Rikki Schlott call an "adulthood of the American mind."[54]

Regulating Social Media

Of course, there is a darker and less free future as well. Government regulation that requires or bans certain moderation techniques and decisions could severely harm the online experience of social media users while chilling innovation and new online experiences. There are multiple ways in which this could go wrong. In Europe, existing hate speech laws are regularly used to silence speech both online and offline, including all sorts of political and social commentary, such as:

- blasphemy laws in Spain that have been used to prosecute woman's rights activists;[55]
- German laws against Nazi symbols that were used to convict a German father who compared the actions of a government employment agency to those of the Nazi in discriminating against his mixed-race daughter;[56]
- a Danish sacrilege law that was passed in 2023 to stop the burning of Qurans;[57] and
- the 2016 conviction of Netherland's far-right MP Geert Wilders—whose political party won a plurality of

votes in 2023—for a campaign speech asking if voters wanted more Moroccans in the Netherlands.[58]

As Nadine Strossen lays out in *HATE: Why We Should Resist It with Free Speech Not Censorship*, hate speech laws are unnecessarily broad and vague, which ensures that they inevitably silence even those people that they are meant to protect.[59] The new European Digital Services Act, for example, has already been abused by censors in Brussels to bring formal investigations against major platforms and actual prosecution of X (formerly Twitter) for allowing "disinformation" and "illegal content," particularly regarding the conflict in Israel and Gaza.[60] The EU's conflation of "disinformation" and "illegal content," as well as its arbitrary demands for responses in a 24-hour timeframe to its missives to social media companies following the October 7 attack on Israel, represent an effort to pressure companies to remove disfavored speech—alluded to as anti-Israeli or anti-Semitic speech—surrounding an active conflict, an effort that goes beyond the already expansive Digital Services Act regulations.[61] Tech companies increasingly face the choice about whether to widely suppress content that EU bureaucrats dislike or pay up to 6 percent of their global revenue in fines. Requiring tech companies to comply with European or other nations' laws governing online expression all but guarantees that innovations that give users more control and a better experience will be limited.[62]

"No matter the rationale, government regulations will likely harm the expression and experiences of users online."

On the other hand, regulations that force platforms to carry speech are similarly problematic. As mentioned earlier, viewpoint neutrality requirements will spawn endless questions about what is and what is not a viewpoint and whether a social media company can prove its neutrality.[63] Thus, viewpoint neutrality requirements could result in anything resembling a controversial viewpoint either being banned or allowing even the most heinous of viewpoints to avoid legal liability. Even requirements to host politicians' speech create a large loophole in a platform's content policies, meaning every troll can break the rules as long as they've declared a run for political office.[64] Other regulations that require social media companies to host or carry speech could chill the development of new moderation services by limiting what those services are legally able to provide to users. They could also shut down online services that seek to specifically serve particular communities or viewpoints.

No matter the rationale, government regulations will likely harm the expression and experiences of users online.

CONCLUSION

The essentials of content moderation should inform policymakers as they think about content regulation. At its core, content policies are rules that organizations use to create their preferred spaces. As private property, these platforms have decided to create rules and use moderation techniques to turn their property into a virtual space that aligns with their interests. Just as a country club may set rules for its members to create the kind of environment it wants, in the same way, tech companies that host user-generated content may craft rules that determine what their online environment will be. For government to compel them to host users and speech contrary to those rules would abridge their speech and property rights.[65] This may frustrate some policymakers who view various aspects of content moderation as biased one way or another. But a platform's bias, even if explicitly stated, is no different than a newspaper choosing what op-eds make it into the paper or a bookseller choosing which books to put in the shop window.

Beyond these important free expression realities, government regulation of content moderation also has practical challenges posed by moderating content at scale. No matter the animating principles, platforms need policies they can actually implement. Policymakers may want to force companies to moderate according to certain principles or to stop certain kinds of speech, but such goals may be unworkable in practice and may even backfire against the government's objectives. In the long run, government regulations will likely chill the development of new online services or means of moderation that may empower and give more choices to users, erect barriers to new businesses, and hinder a culture of free expression.

Rather than government force, the market is already adapting and providing users with the experiences and tools that fit their demands. Yes, social media companies can be—and are—imperfect and biased. Many claim, however, to value input from diverse perspectives as they develop their policies. Unfortunately, many who want greater free expression often do not engage with these companies, and so they have little direct input into policy or product development. On the other hand, viewpoints and interest groups that are more hostile to free expression dominate the content-moderation space. Complaining about ad hoc policy moderation actions doesn't address this lack of liberty-minded civil society engagement with the underlying policies themselves. As can be seen in countless scenes across our society, we cannot assume that everyone understands or agrees with the importance of free expression. Markets and civil society, rather than government regulation, are essential if we are to see a better future online.

NOTES

1. In many ways, I am personally frustrated with the state of content moderation at some of the largest social media companies, not just as a user, but as a former member of Meta's content policy team responsible for the community standards. But despite free expression concerns that I may have with the state of content moderation, that does not change the facts about how content is moderated, the consequences of government intervention, and the right of private organizations to run their services as they choose.

2. "Meta Reports Third Quarter 2023 Results," Meta Platforms Inc., press release, October 25, 2023.

3. "Community Guidelines Enforcement Report: April 1, 2023–June 30, 2023," TikTok, October 2023. Author calculations are based on the absolute number of removals and the removal rate of published content. This fits in between other published estimates.

4. "YouTube for Press," *YouTube Official Blog* (blog).

5. Eugene Volokh, "*Volokh v. N.Y. A.G.*: "New York Can't Target Protected Online Speech by Calling It 'Hateful Conduct,'" *Reason*, December 1, 2022.

6. "Section 230: An Overview," Congressional Research Service, R46751, January 4, 2024; and "Section 230," Electronic Frontier Foundation.

7. *303 Creative LLC v. Elenis*, 143 S. Ct. 2298 (2023).

8. Eli Lehrer, "Public Spaces in the Digital Age," R Street Institute, September 20, 2021.

9. "Policies and Guidelines," YouTube Creators; "Policies," Meta Transparency Center; and "Rules and Policies," X Help Center.

10. "Terms of Use," Facebook via Wayback Machine, June 28, 2005.

11. This does not include the copyright sections, which add about another 150 words to both the original and current policies. This also does not include additional policies that go beyond the current community standards for pages, groups and events, recommended content, advertising, commerce, etc.

12. Yoel Roth, "Content Moderation's Legalism Problem," *Lawfare*, July 24, 2023.

13. Solána Imani Rowe, "SZA—Kill Bill Music Video," SZA, posted January 10, 2023, YouTube video, 4:35.

14. "Policy Forum Minutes," Meta Transparency Center, updated December 27, 2023; "Our Approach to Policy Development and Enforcement Philosophy," X Help Center; and "Community Guidelines," YouTube Rules and Policies.

15. Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou, *Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies* (Copenhagen: Future of Free Speech, July 2023); Eric Hananoki, "As Musk Endorses Antisemitic Conspiracy Theory, X Has Been Placing Ads for Apple, Bravo, IBM, Oracle, and Xfinity Next to Pro-Nazi Content," Media Matters, November 16, 2023; "Six Things ADL Is Watching Following Meta's Threads Launch," *ADL* [Anti-Defamation League] (blog), July 13, 2023; "Stop Hate for Profit," Stop Hate for Profit; and "GLAAD's Third Annual Social Media Safety Index Shows All Five Major Social Media Platforms Fail on LGBTQ Safety," GLAAD, June 15, 2023.

16. "Where We Have Fact Checking," Meta for Media.

17. Greg Lukianoff and Rikki Schlott, *The Canceling of the American Mind* (New York: Simon & Schuster, 2023), pp. 51–62.

18. Eric Kaufmann, "Academic Freedom in Crisis: Punishment, Political Discrimination, and Self-Censorship,"

Center for the Study of Partisanship and Ideology, March 1, 2021; and Greg Lukianoff and Rikki Schlott, *The Canceling of the American Mind* (New York: Simon & Schuster, 2023), pp. 36–37, 56–58.

19. Sacha Altay et al., "A Survey of Expert Views on Misinformation: Definitions, Determinants, Solutions, and Future of the Field," *Misinformation Review*, Harvard Kennedy School, July 27, 2023; and "Policy Forum Minutes," Meta Transparency Center, updated December 27, 2023. For example, see this list of examples that Meta provides on how it engages with external parties in its policy development process. The academics that are specifically mentioned are working on hate speech, social psychology, history, journalism and communications, and law. "Input from External Stakeholders," Meta Transparency Center, updated January 18, 2023.

20. David Inserra, "The Free Speech Recession Deepens across the Democratic World," *Cato at Liberty* (blog), Cato Institute, January 8, 2024; and Meri Baghdasaryan and Karen Gullo, "UN Human Rights Committee Criticizes Germany's NetzDG for Letting Social Media Platforms Police Online Speech," *Electronic Frontier Foundation*, November 23, 2021.

21. Andrew Grossman and Kristin A. Shapiro, "Shining a Light on Censorship: How Transparency Can Curtail Government Social Media Censorship and More," Cato Institute Briefing Paper no. 168, October 3, 2023; and David Inserra, "Lawsuit Alleges More Government Censorship by Proxy—State Department Funds Blacklisting of US Media?," *Cato at Liberty* (blog), Cato Institute, December 13, 2023.

22. John Bowden, "Twitter CEO Jack Dorsey: I 'Fully Admit' Our Bias Is 'More Left-Leaning,'" *The Hill*, August 18, 2018; and Blair Guild, "Sen. Ted Cruz Grills Mark Zuckerberg about Facebook Political Bias," *CBS News*, April 10, 2018.

23. "Alphabet Inc.," OpenSecrets, February 2, 2024; "Netflix Inc.," OpenSecrets, 2024; "Twitter," OpenSecrets, February 2, 2024; and "Meta," OpenSecrets, February 2, 2024.

24. "Reddit Content Policy," Reddit.

25. "What Is Mastodon?," Mastodon, last updated January 14, 2024; and Amanda Silberling and Alyssa Stringer, "What Is Bluesky? Everything to Know about the App Trying to Replace Twitter," *TechCrunch*, November 17, 2023.

26. "About Fediverse," Fediverse, last updated January 1, 2024; and Mike Solana, "The End of Social Media: An Interview with Jack Dorsey," Pirate Wires, May 9, 2024.

27. "An Update on Twitter Transparency Reporting," *X Blog* (blog), April 25, 2023.

28. "Community Standards Enforcement Report," Meta, November 2023.

29. "The Comments Section," Help, *New York Times*; and "Detecting Violations," Meta Transparency Center.

30. Jonathan Vanian, "TikTok Has Tens of Thousands of Moderators Led by Group in Ireland Looking for Offensive Content, CEO Says," *CNBC*, updated April 21, 2023.

31. TikTok, "Community Guidelines Enforcement Report," October 2023.

32. "Reddit Content Policy," Reddit.

33. "AutoModerator," Reddit; and "Moderator Code of Conduct," Reddit, September 25, 2023.

34. Bluesky Team, "Moderation in a Public Commons," *Bluesky Blog* (blog), June 23, 2023; and "Bluesky's Stackable Approach to Moderation," Bluesky, March 12, 2024.

35. Mkantzer, "0002 Labeling and Moderation Controls," GitHub, last updated June 24, 2023.

36. "My Account Was Banned for Violating Reddit's Rules," Reddit Help Center, Reddit, last updated September 2023; "Appealed Content," Meta Transparency Center, last updated November 18, 2022; and "Account Safety," Help Center, TikTok.

37. David McCabe, "Evidence of Anti-Conservative Bias by Platforms Remains Anecdotal," *New York Times*, October 28, 2020; Heather Moon and Gabriela Pariseau, "Alarming Election Interference! Big Tech Censors Biden Opponents 162 Times," Media Research Center, November 28, 2023; and Steven Overly and Alexandra S. Levine, "Trump Draws Public into Bias Feud with Social Media Firms," *Politico*, May 15, 2019.

38. "Do Not Post Violent Content," Help Center, Reddit, last updated August 2023.

39. "Facebook Community Standards: Hate Speech," Meta Transparency Center, last updated January 30, 2024; and Rob Picheta, "Instagram and Facebook Ban All Content Promoting Conversion Therapy," *CNN Business*, July 11, 2020. The policy claims to remove any services designed to change one's sexual orientation or gender identity. However, based on news reports of the announced policy, this policy only prohibits so-called "conversion therapy" efforts to promote heterosexual or a cisgender identities.

40. Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou, *Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies* (Copenhagen: Future of Free Speech Project, July 2023).

41. Nadine Strossen, *HATE: Why We Should Resist It with Free Speech, Not Censorship* (Oxford: Oxford University Press, 2018).

42. Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou, *Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies* (Copenhagen: Future of Free Speech Project, July 2023).

43. Thomas A. Berry, "New York's Hate Speech Law Violates the First Amendment," *Cato at Liberty* (blog), Cato Institute, September 23, 2023.

44. Caitlin O'Kane, "Representative Jim Banks Suspended from Twitter after Misgendering Trans Health Official Dr. Rachel Levine," *CBS News*, October 23, 2021; Io Dodds, "Conservative Satire Site The Babylon Bee Locked Out of Twitter for Misgendering Trans White House Official," *Independent*, March 22, 2022; Jamie Burton, "Jordan Peterson 'Would Rather Die' Than Delete Elliot Page Tweet," *Newsweek*, July 4, 2022; Nick Statt, "Facebook and Instagram Ban All Posts Promoting Conversion Therapy," *The Verge*, July 10, 2020; Morgan Sung, "Discord Bans Misgendering and Deadnaming in Hateful Conduct Policy Update," TechCrunch, December 13, 2023; Cristiano Lima, "Facebook No Longer Treating 'Man-Made' Covid as a Crackpot Idea," *Politico*, May 26, 2021; Matt Taibbi, "The British Medical Journal Story That Exposed Politicized 'Fact-Checking,'" *Racket News*, February 1, 2022; Jack Houghton, "EXCLUSIVE: Bombshell Dossier of SECRET Emails Proves Mark Zuckerberg's Facebook Fact Checking Program Has Been Compromised by Activists," March 6, 2024; Elizabeth Culliford, "Facebook Removes Anti-Quarantine Protest Events in Some US States," Reuters, April 20, 2020; and Amy Judd, "GoFundMe for Canada's Trucker Convoy Removed for Violating 'Terms of Service,'" Global News, February 4, 2022.

45. The NFX Team, "The Network Effect's Manual: 16 Network Effects (and Counting), NFX, June 2021; and Bobby Allyn, "Why Can't Twitter and TikTok Be Easily Replaced? Something Called 'Network Effects,'" NPR, April 12, 2023.

46. Catherine Tucker, "What Have We Learned in the Last Decade? Network Effects and Market Power," *Antitrust* (Spring 2018): 77–81; Marco Iansiti, "Assessing the Strength of Network Effects in Social Network Platforms," Harvard Business School Working Paper no. 21-086, 2021; and Jonathan A. Knee, "Network Effects Are Overrated," *New York Times*, updated September 6, 2021.

47. David Thiel et al., *Contours and Controversies of Parler* (Stanford, CA: Internet Observatory, Cyber Policy Center, January 2021).

48. Gabe Bullard, "Six Months Ago NPR Left Twitter. The Effects Have Been Negligible," *NiemanReports*, October 11, 2023.

49. Roxanna Woloshyn and Ashley Fraser, "What Is the Fediverse and Why Does Threads Want to Join?," *CBC News*, July 19, 2023; Mike Masnick, "Protocols, Not Platforms: A Technological Approach to Free Speech," Knight First Amendment Institute, Columbia University, August 21, 2019; Adam Zewe, "Empowering Social Media Users to Assess Content Helps Fight Misinformation," *MIT News*, November 16, 2022; and Francis Fukuyama et al., *Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy* (Stanford, CA: Cyber Policy Center, Freeman Spogli Institute).

50. Thor Benson, "The Small but Mighty Danger of Echo Chamber Extremism," *WIRED*, January 20, 2023; and NPR Staff, "The Reason Your Feed Became an Echo Chamber—and What to Do about It," NPR, July 24, 2016.

51. Naijia Liu et al., "Algorithmic Recommendations Have Limited Effects on Polarization: A Naturalistic Experiment on YouTube," September 18, 2023; and Andrew M. Guess et al., "How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?," *Science* 381, no. 6656 (July 2023): 398–404.

52. Greg Lukianoff and Rikki Schlott, *The Canceling of the American Mind* (New York: Simon & Schuster, 2023), pp. 188–90.

53. Jacob Mchangama, *Free Speech: A History from Socrates to Social Media* (New York: Basic Books, 2022), p. 5.

54. Greg Lukianoff and Rikki Schlott, *The Canceling of the American Mind* (New York: Simon & Schuster, 2023), pp. 304–5.

55. "Spain," End Blasphemy Laws.

56. Jacob Mchangama, *Free Speech: A History from Socrates to Social Media* (New York: Basic Books, 2022), p. 346.

57. David Inserra, "The Return of Blasphemy and Sacrilege Laws in the Most Unlikely Places," *Cato at Liberty* (blog), Cato Institute, January 23, 2024.

58. Nadine Strossen, *HATE: Why We Should Resist It with*

Free Speech, Not Censorship (Oxford: Oxford University Press, 2018), pp. 97.

59. Nadine Strossen, *HATE: Why We Should Resist It with Free Speech, Not Censorship* (Oxford: Oxford University Press, 2018).

60. "Europe: Tackling Content about Gaza and Israel Must Respect Rule of Law," *Article 19*, October 18, 2023.

61. Thierry Breton (@Thierry Breton), "Following the terrorist attacks by Hamas against [Israel], we have indications of X/Twitter being used to disseminate illegal content & disinformation in the EU," X post, October 10, 2023, 2:20 p.m.

62. J. D. Tuccille, "EU's Digital Services Act Threatens Americans' Free Speech," *Reason*, June 5, 2023.

63. H.B. 20, 87th Gen. Assemb., 1st C.S., (Tx. 2021).

64. S.B. 7072, 53rd Sess., Reg. Sess., (Fl. 2021).

65. Thomas A. Berry and Anastasia P. Boden, *NetChoice v. Paxton & Moody v. NetChoice*, Cato Institute Legal Briefs, December 7, 2023.

RELATED PUBLICATIONS FROM THE CATO INSTITUTE

Courts Should Affirm First Amendment Rights of Youths in the Digital Age: The Case for a 21st–Century *Tinker* by Jennifer Huddleston, Briefing Paper no. 176 (March 28, 2024)

NetChoice v. Paxton & Moody v. NetChoice by Thomas A. Berry and Anastasia P. Boden, Legal Briefs (December 7, 2023)

Shining a Light on Censorship: How Transparency Can Curtail Government Social Media Censorship and More by Andrew M. Grossman and Kristin A. Shapiro, Briefing Paper no. 168 (October 3, 2023)

A Link Tax Won't Save the Newspaper Industry: The Journalism Competition and Preservation Act Will neither Promote Competition nor Preserve Newspapers by Paul Matzko, Policy Analysis no. 956 (August 14, 2023)

Competition and Content Moderation: How Section 230 Enables Increased Tech Marketplace Entry by Jennifer Huddleston, Policy Analysis no. 922 (January 31, 2022)

RECENT STUDIES IN THE CATO INSTITUTE POLICY ANALYSIS SERIES

973. **A Case for Federal Deficit Reduction Spending Cuts to Avoid a Fiscal Crisis** by Ryan Bourne (April 18, 2024)

972. **Terrorism and Immigration: A Risk Analysis, 1975–2023** by Alex Nowrasteh (April 9, 2024)

971. **A Return to US Casualty Aversion: The 9/11 Wars as Aberrations** by John Mueller (April 2, 2024)

970. **Biden Short-Term Health Plans Rule Creates Gaps in Coverage: Rule Would Deny Care after Patients Fall Ill** by Michael F. Cannon (March 14, 2024)

969. **State Fiscal Health and Cost-Saving Strategies** by Marc Joffe (February 20, 2024)

968. **Bo/d International Tax Reforms to Counteract the OECD Global Tax** by Adam N. Michel (February 13, 2024)

967. **Containing Medicaid Costs at the State Level** by Marc Joffe and Krit Chanwong (February 6, 2024)

966. **Curbing Federal Emergency Spending Government Spending Grows with Excessive and Wasteful Emergency Designations** by Romina Boccia and Dominik Lett (January 9, 2024)

965. **Taiwan's Urgent Need for Asymmetric Defense** by Eric Gomez (November 14, 2023)

964. **Trade and Investment Are Not a Balancing Act** by Norbert J. Michel (November 7, 2023)

963. **Misperceptions of OPEC Capability and Behavior: Unmasking OPEC Theater** by David Kemp and Peter Van Doren (November 2, 2023)

962. **Are Public School Libraries Accomplishing Their Mission? Public School Libraries Do Not Appear to Stock a Balance of Views** by Neal McCluskey (October 17, 2023)

961. **Pariah or Partner? Reevaluating the US-Saudi Relationship** by Jon Hoffman (September 20, 2023)

960. **Expand Access to Methadone Treatment: Remove Barriers to Primary Care Practitioners Prescribing Methadone** by Jeffrey A. Singer and Sofia Hamilton (September 7, 2023)

959. **Sweden during the Pandemic: Pariah or Paragon?** by Johan Norberg (August 29, 2023)

958. **Terrorism and Immigration: A Risk Analysis, 1975–2022** by Alex Nowrasteh (August 22, 2023)

957. **Corking Russian Gas: Global Economic and Political Ramifications** by Scott Semet (August 17, 2023)

956. **A Link Tax Won't Save the Newspaper Industry: The Journalism Competition and Preservation Act Will neither Promote Competition nor Preserve Newspapers** by Paul Matzko (August 14, 2023)

955. **Freeing American Families: Reforms to Make Family Life Easier and More Affordable** by Vanessa Brown Calder and Chelsea Follett (August 10, 2023)

954. **Tax Expenditures and Tax Reform** by Chris Edwards (July 25, 2023)

953. **2022 Arms Sales Risk Index** by Jordan Cohen and A. Trevor Thrall (July 18, 2023)

952. **Adverse Effects of Automatic Cost-of-Living Adjustments to Entitlement and Other Payments** by John F. Early (June 22, 2023)

951. **Indian Nationalism and the Historical Fantasy of a Golden Hindu Period** by Swaminathan S. Anklesaria Aiyar (June 21, 2023)

950. **Why Legal Immigration Is Nearly Impossible: US Legal Immigration Rules Explained** by David J. Bier (June 13, 2023)

949. **Global Inequality in Well-Being Has Decreased across Many Dimensions: Introducing the Inequality of Human Progress Index** by Chelsea Follett and Vincent Geloso (June 8, 2023)

948. **The High Price of Buying American: The Harms of Domestic Content Mandates** by James Bacchus (June 6, 2023)

947. **The Future of the WTO: Multilateral or Plurilateral?** by James Bacchus (May 25, 2023)

946. **Course Correction: Charting a More Effective Approach to US-China Trade** by Clark Packard and Scott Lincicome (May 9, 2023)

945. **The Right to Financial Privacy: Crafting a Better Framework for Financial Privacy in the Digital Age** by Nicholas Anthony (May 2, 2023)

944. **Balance of Trade, Balance of Power: How the Trade Deficit Reflects US Influence in the World** by Daniel Griswold and Andreas Freytag (April 25, 2023)

943. **Streamlining to End Immigration Backlogs** by David J. Bier (April 20, 2023)

942. **Transforming the Internal Revenue Service** by Joseph Bishop-Henchman (April 11, 2023)

CITATION

Inserra, David. "A Guide to Content Moderation for Policymakers," Policy Analysis no. 974, Cato Institute, Washington, DC, May 21, 2024.