TRUST & TRADE-OFFS //

Pathways to Effective Platform Transparency & Moderation

Amy Benziger and Chelsea Magnant

September 2022

CITRIS AND THE BANATAO INSTITUTE | CITRIS POLICY LAB

OUR BETTER WEB

EXECUTIVE SUMMARY

There is growing public concern about harmful online content and a corresponding call for regulation of platforms, with a specific focus on Section 230 of the Communications Decency Act. While these reforms are grounded in the mission to create healthier online environments, there has not been enough conversation on *what* meaningful indicators of what "healthier"look like and *why* identifying the right technology and policy strategies to achieve that vision are so challenging.

The purpose of this project is twofold:

1. Provide an overview of the current content moderation landscape and efforts to reform it to better ensure a greater diversity of stakeholders have a common understanding of the challenges.
2. Help the reader think about the future of content reform in a more nuanced, analytic, and future-focused way. Implementing new rules requires trade-offs, and it's important to think through the positive and negative effects of rebalancing those trade-offs.

The paper begins with a challenge to the reader to put themselves into the mindset of a digital citizen, a government official, or an in-house content moderator. What we mean by a digital citizen, rather than simply a digital consumer, implies confronting the real-world trade-offs that are unavoidable in a system that operates on a global scale across many political and societal contexts. We then lay out general principles we used to guide our analysis throughout the paper. While many have debated elements of these principles, our goal in outlining them upfront is to establish the lens through which to read the paper.

The following section gives a general overview of challenges inherent to the current moderation landscape. It is followed by a summary of the legislation targeting Section 230 from the 116th and 117th Congressional sessions. We offer our analysis of legislative themes, as well as a brief outline of the political perspectives surrounding the content moderation conversation.

We detail why proposed legislation is unlikely to work, or be passed at all. To do that, we explore:

1. the benefits and challenges of the First Amendment and how it uniquely influences America's relationship to our digital landscape,
2. the massive scale of content moderation and the intrinsic inevitability of conflict, and
3. the lack of data to universally define societal harm in regards to a digital action.

Lastly, we encourage the reader to explore four scenarios to help them think through what their ideal version of healthier platforms look like and what trade-offs it might take to arrive there. The paper concludes with a series of recommendations.

SUMMARY OF RECOMMENDATIONS

We believe that at a baseline, healthier platforms are more collaborative and trust-based. Given that, we propose that the calls we hear for "transparency" from all sides—academia, civil society, government, and industry—actually reflect a desire for a mutual sense of trust. Trust in this context does not mean a sense of blind faith. It means a belief in a system of checks and balances among all parties that protects against unconstrained and unchecked power. To move toward this type of a system, we detail four recommendations that can be implemented to support effective transparency and moderation.

1. Maximize User Control: Support greater portability, interoperability, and delegability allowing users to move their own data between platforms, easily communicate with people on different operating systems, and enable users to assign ownership to third-party systems to manage the flow of information on the platforms they choose to use.

2. Use Data to Inform Research and Oversight: Empower access for a trusted group of qualified researchers—academics, civil society organizations, and journalists—to secure data from social media platforms with the intention of creating an oversight mechanism on the design and implementation of the systems, rather than the content of the speech.[1] The goal is independent verification and pressure on how these products impact social outcomes and what role the government should play in ensuring the protection of public safety.

3. Focus on "Friction": Experiment with and document the effects of adding points of friction into social media platforms. Friction, in this case, means adding in tactics to slow down the uncontrolled virality of posts with the intention of ensuring that the fast spread is in line with each platform's moderation practices.

4. Educate the Next Generation: Incorporate digital literacy curriculum into US schools to educate our youth to become responsible digital citizens.

[1] Frances Haugen. "Europe Is Making Social Media Better without Curtailing Free Speech. the U.S. Should, Too." The New York Times, April 28, 2022. https://www.nytimes.com/2022/04/28/opinion/social-media-facebook-transparency.html?referringSource=articleShare.

INTRODUCTION

Congress is currently considering dozens of bills to amend, repeal, or reform Section 230, yet most are unlikely to pass. Others face steep hurdles. Media outlets and academics are intensifying their scrutiny of online platforms with few resulting changes. Despite the growing level of public discontent with platforms and the plethora of harmful content on them, progress is stagnant. Why?

We started our research project with a lofty goal in mind: to offer a new policy recommendation for platform governance. We conducted more than 30 interviews with experts from platforms, nonprofits, the federal government, and leading research universities. We reviewed hundreds of blog posts, academic journals, press articles, and podcasts about content reform. And we encountered numerous, varying, and equally valid opinions about Section 230 and how the state of public online discourse is changing our society, for better and for worse.

However, we found very few thorough explanations of *why* our content problems are hard to solve. And without a comprehensive understanding of what's happening in this space—including the scale and scope of harmful content, the dynamics within tech companies, and the existing legal framework through which we view this problem—we believe society will continue the debate without a clear and satisfactory resolution.

In this paper, we focus on two main questions:

1. Why has it been so difficult to pass any of the proposed legislation targeting Section 230?
2. What should we do to mitigate harmful content online?

What is Section 230?

Section 230 of the Communications Decency Act was passed in 1996 with the goal of protecting the emerging Internet landscape in the United States. There had been a series of libel suits against Internet service providers (ISPs) for defamatory content on their websites, and Section 230 authors Ron Wyden (D-OR) and Chris Cox (R-CA) saw this as a major impediment to growth.

The Act states, in part, that "no provider or user of an interactive computer service shall be treated as a publisher or speaker of any information provided by another content provider."[2] This became the legal and policy framework that protects websites from civil liability should a user post something offensive or illegal on their platforms, with exceptions for copyright violations, sex work-related material, and violations of federal criminal law. It also provides immunity for the platforms to moderate content based on their own discretion and platform rules.

It is simultaneously one of the Internet's greatest tools to protect freedom of speech and innovation *and* what has been blamed for the proliferation of hate speech, misinformation, extremism, and other harmful content.

[2] "47 U.S. Code § 230 - Protection for Private Blocking and Screening of Offensive Material." Legal Information Institute. Accessed May 5, 2022. https://www.law.cornell.edu/uscode/text/47/230.

THE CHALLENGE TO THE READER

Many people say we need greater transparency to hold online platforms accountable for the content they host, but we have yet to clearly and universally agree on what that entails. We believe what people really want when they say “transparency” is trust in a system of checks and balances. In order to create healthier platforms, we must create a trust-based system where all parties—academia, civil society, government, the platforms, and users—trust each other with the data they are providing and the accountability surrounding that data.

The initial launch of web services, especially social media platforms, sought to bring people together to disseminate new and different ideas on a global scale. While those core elements of connectivity still exist, the question is how to ensure that those points of connectivity are primarily collaborative, not combative. A healthier platform is not just a better interface; it’s a better relationship among all those with a stake in the system.

It’s easy to call for action on the sidelines; it is harder to confront trade-offs that are inevitable in a system fueled by the real-time content of billions of individuals with varying cultures, languages, viewpoints, and motivations. While reading this paper, we encourage our readers to put themselves into the mindset of a digital citizen, a government official, or an in-house content moderator. What trade-offs might you be willing to make as you envision your version of a healthier platform?

Roles & Questions for the Reader

As a **digital citizen**, what would give you a greater sense of transparency than what the platforms currently provide?

As a **government official**, what levers could and would you use to try to achieve a more transparent system? How would you enforce those rules without infringing on user or platform rights?

As an **in-house content moderator**, how would you mitigate the inherent tension between the calls for transparency, security, and privacy?

As you consider these questions, we also challenge you to think beyond the platforms that frequently dominate the news cycle—especially Facebook—to any platform that utilizes user-generated content: 4chan, Pinterest, Twitch, Wikipedia, or Yelp.

We hope to spark discussion and help readers think more critically about issues that have been cast as binary. This is neither an apology for platform missteps, nor a blank check for governments or users. It’s intended in the spirit of finding space to build trust surrounding those tensions. It has become clear to all parties that real change—whether technological or regulatory—can and should happen.

GENERAL PRINCIPLES

The following five principles should be used to guide the reader:

1. Platforms Have First Amendment Rights

The First Amendment applies only to government suppression of speech. While states like Florida and Texas have passed laws that aim to prohibit large social media companies from banning users or removing content based on political viewpoints, the constitutionality of these laws has come under scrutiny.[3] Platforms have the right as private entities to deny any kind of speech that doesn't align with the content moderation rules they've created.

2. There is Too Much Content to Moderate Perfectly in Real-Time

As of February 2020, more than 30,000 hours of video were uploaded to YouTube and roughly 350,000 tweets and 510,000 Facebook comments were posted every hour.[4] Across Facebook, Messenger, Instagram, and WhatsApp, 1 billion stories are shared around the world every day.[5] Moderating this amount of content at scale and in real-time becomes even more complex when you consider the world's many languages and social contexts that moderators have to navigate when making content decisions. Moreover, content moderation decisions are inherently subjective, and most decisions occur in a gray area influenced by everything from coded language to suggestive imagery. Those who have had their content moderated are most likely going to disagree with the decision to have it taken down. This inherently means there is no way to moderate in a universally agreed upon way. According to a 2020 NYU Stern report, Facebook's moderators review about 3 million posts, photos, and videos that have been flagged by AI or users a day, and Mark Zuckerberg has admitted content moderators make the wrong call more than 10% of the time. That means about 300,000 content mistakes are made *per day* on Facebook alone.[6]

3. Content Moderation is Not Binary

The prevailing narratives are that platforms are either moderating too much or too little. However, that explanation is overly simplistic. Many platforms are trying in earnest to improve moderation practices, *and* they have not devoted enough resources to content moderation, do not give moderators sufficient time to make challenging content decisions that have long-term consequences, and do not properly support outsourced moderation teams.[7] Not enough resources

[3] Cat Zakrzweski, "11th Circuit blocks major provisions of Florida's social media law", Washington Post, May 23, 2022, https://www.washingtonpost.com/technology/2022/05/23/florida-social-media-11th-circuit-decision/. Mark Sullivan. "Why the Texas Social Media Law Just Became a Big Headache for Big Tech," Fast Company, May 17, 2022, https://www.fastcompany.com/90752528/why-the-texas-social-media-law-just-became-a-big-headache-for-big-tech.

[4] "57 Fascinating and Incredible YouTube Statistics." Brandwatch. Accessed May 6, 2022, https://www.brandwatch.com/blog/youtube-stats/ ; "Twitter Usage Statistics." Twitter Usage Statistics - Internet Live Stats. Accessed July 1, 2022, https://www.internetlivestats.com/twitter-statistics/.

[5] "Wild and Interesting Facebook Statistics and Facts (2022)." Kinsta®, January 3, 2021, https://kinsta.com/blog/facebook-statistics/.

[6] Paul Barrett. "Who Moderates the Social Media Giants? A Call to End Outsourcing." Issuu. Accessed May 6, 2022. https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version?fr=sZWZmZjI1NjI1Ng.

[7] Billy Perrigo. "Facebook Faces New Lawsuit Alleging Human Trafficking and Union-Busting in Kenya." Time Magazine, May 11, 2022. https://time.com/6175026/facebook-sama-kenya-lawsuit/.

have been dedicated to correct bias within AI technology, *and* AI technology is not yet able to make the kind of nuanced decisions necessary to effectively moderate content on a global scale. These issues are complex, and the policy creation process to account for that complexity is typically iterative, which can seem unacceptably slow to those outside the platforms. No single solution will satisfy everyone, but we are seeing growing evidence of collaboration among platforms and technology professionals through groups like the Digital Trust and Safety Partnership and the Integrity Institute to create shared standards and practices.[8]

4. Algorithms are Driven by Attention

Algorithms dictate much of how and where our collective attention is focused, and platforms like Instagram or TikTok show you content they think you'll like. Roughly 70% of recommended YouTube videos come from recommendation algorithms that are built on data collected from users.[9] There is an inherent tension in the push and pull of these algorithms. For example, while you may not have searched for content that is served to you, the algorithm learned from content you've previously searched for or interacted with to infer what it should deliver. While platforms may not have full editorial control over all user-generated content, some have the ability to manipulate the lens through which you see your virtual world with their algorithms, which is arguably more powerful than an editorial pen.

5. Not All Platforms are Equal

We talk about platforms as a monolithic entity, but each platform has its own features, size, and scope. Proposed solutions may be effective for one platform, but could make things worse for another. Sweeping changes—like making platforms liable for all automated content moderation decisions—are often economically feasible for big companies like Meta and Twitter but would present huge challenges for small platforms without the resources for automated screening or extensive content moderation teams.

[8] Digital Trust and Safety Partnership. Accessed May 2, 2022, https://dtspartnership.org/; Integrity Institute. Accessed May 2, 2022, https://integrityinstitute.org/.

[9] Paige Cooper. "How Does the YouTube Algorithm Work in 2021? The Complete Guide," Social Media Marketing & Management Dashboard." Hootsuite Blog, June 21, 2021. https://blog.hootsuite.com/how-the-youtube-algorithm-works/#:~:text=The%20goal%20was%20to%20find,watching%20videos%20the%20algorithm%20recommends.

THE CURRENT MODERATION LANDSCAPE

Platform moderation refers to the efforts of in-house policy teams, (mostly) outsourced moderation teams, AI, and internal content moderation policies to limit the amount of harmful and illegal content online. The thresholds for different kinds of content vary by platform, except for those issues that are illegal regardless of medium (e.g., child sexual abuse material). While content moderation began as a way to create safer spaces on the Internet, many people now think content moderation teams should identify misinformation, safeguard democracy, and prevent social unrest, all in real-time.

It is often difficult for content moderators to decide if content has violated terms of use and/or community standards. One clear example of this is what many "trust and safety" professionals—those who develop and enforce the policies defining acceptable behavior and content online—described as "harmful and dangerous" content.[10] This category includes the viral challenges and threatening pranks that periodically become popular, such as the "Tide Pod Challenge."[11] These tech employees are responsible for trying to stay ahead of every bizarre new trend that arises, and each one we interviewed—regardless of platform—talked about the constant stress they feel as they learn about new facets of what their jobs entail. Were they supposed to have predicted teenagers would start eating Tide Pods? Where is the line between what one could consider a minor prank and what could inspire others to self-harm? How does a new, emerging trend rank among the many others they're monitoring at the same time?

Moderators face similar struggles with coded language. Many extremist and conspiracy groups regularly change keywords and visuals to avoid detection, which keeps moderators guessing about what content is inciting violent or harmful behavior. Qanon is perhaps the most well-known example; social networks amended policies to limit Qanon's ability to organize, prompting the group to adopt visuals, memes, screenshots, and coded language to evade detection.[12] Similarly, teenagers are using emojis and coded language to find and buy drugs.[13]

Although we understand the challenges faced by the policy developers and enforcers we spoke to, we also acknowledge that some platforms are not dedicating enough resources to these moderation challenges. Let's take Facebook for example. The "Facebook Files" leaked in 2021 depict a company struggling to contend with, and often neglecting, known content challenges, particularly in languages other than English. The files revealed that 87% of Facebook's operational budget to combat disinformation goes to protecting users in the US, who make up just 10 % of its

[10] Trust & Safety Professional Association. Accessed April 12, 2022. https://www.tspa.org/.

[11] Lindsey Bever. "Teens are daring each other to eat Tide pods. We don't need to tell you that's a bad idea." The Washington Post, January 17, 2018. https://www.washingtonpost.com/news/to-your-health/wp/2018/01/13/teens-are-daring-each-other-to-eat-tide-pods-we-dont-need-to-tell-you-thats-a-bad-idea/.

[12] Lydia Morrish. "How Qanon Content Endures on Social Media through Visuals and Code Words." First Draft, March 4, 2021. https://firstdraftnews.org/articles/how-qanon-content-endures-on-social-media-through-visuals-and-code-words/.

[13] Elise Sole. "Experts Are Warning Parents about a Popular Shorthand Teens Are Using to Make Online Drug Deals." TODAY, June 14, 2022. https://www.today.com/parents/parents/teens-emojis-online-lingo-drug-deals-rcna33350.

user base.[14] These documents show Facebook employees internally questioned the efficacy of the company's moderation technology and approach, yet Facebook did not appear to respond to many of the concerns until the papers were publicly released and widespread scrutiny followed.[15]

One of the many reasons that Facebook and so many other platforms want to impact how, if at all, Section 230 is changed is to avoid impacts on their business models. To date, the law has afforded these platforms legal protection to experiment with different forms of enforcement in an iterative learning process. However, due to increased scrutiny by both the public and by policymakers, that law and its protection is now up for debate.

[14] Frances Haugen. "Civil society must be part of the Digital Services Act." The Financial Times, March 29,2022. https://www.ft.com/content/99bb6c10-bb09-40c0-bdd9-5b74224a5086.

[15] Tom Simonite. "Facebook Is Everywhere; Its Moderation Is Nowhere Close." Wired, October 25, 2021. https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/.; Gilad Edelman, "How to Fix Facebook, According to Facebook Employees," *Wired*, October 25, 2021, https://www.wired.com/story/how-to-fix-facebook-according-to-facebook-employees/.

AN EXPLORATION OF LEGISLATION TARGETING SECTION 230

Below is a broad overview of how policymakers attempting to reform Section 230 have defined the problem and potential solutions to-date, followed by our analysis of the hurdles facing each bill.[16] For an in-depth review, the CITRIS Policy Lab has created a database of all federally proposed legislation from the 116th and 117th Congressional sessions that explicitly reference Section 230.[17] Many other pieces of legislation informed our research, some of which are noted in this paper. However, the database is limited to this narrow criterion to get a clear understanding of the calls to action by Democrats, Republicans, and bipartisan co-sponsors who aim to directly affect the future of Section 230. Both sides of the aisle have called for reform, and their proposals generally fall under eight themes.

Eight Legislative Themes of Section 230 Legislation

- Complete repeal of Section 230 immunity
- Establish new governing body to oversee technology regulation
- Government-regulated transparency reporting
- Remove Section 230 immunity for over-moderation of content
- Remove Section 230 immunity for under-moderation of content
- Remove Section 230 immunity for algorithmically amplified content
- Remove Section 230 immunity and reclassify large Internet service providers as common carriers
- Remove Section 230 immunity for individual carve-outs of unlawful subject matter such as child sexual abuse material

The primary challenge facing the majority of the proposed bills is that they are not written with clear implementation and accountability mechanisms. Broadly speaking, legislation proposed by Democrats have framed the platforms as valuing profits over people. They want to expand the language of liability to ensure that platforms are held accountable for anything that causes "irreparable harm and/or public risk."[18]

Republican-proposed legislation, on the other hand, wants to narrow the language of immunity from the current framing of "otherwise objectionable" and replace it with concrete terms of "unlawful" content, therefore removing the ability to moderate based on the platform's subjective

[16] The views and findings in this section—and the paper more broadly—are based on research conducted through the CITRIS Policy Lab and are not affiliated with the authors' employers or any other organization.

[17] CITRIS Policy Lab. "Section 230 Legislation Database," Accessed May 27, 2022, https://citrispolicylab.org/section230/.

[18] "47 U.S. Code § 230 - Protection for Private Blocking and Screening of Offensive Material." Legal Information Institute. Accessed May 5, 2022. https://www.law.cornell.edu/uscode/text/47/230.

perception of public harm. Both sides call for increased transparency measures and government oversight, reflecting society's and governments' growing distrust of Internet platforms.

Two broad themes emerged in these congressional sessions, which reflect more nuanced thinking about tech platforms than we saw in the early days of Big Tech-focused congressional hearings:

- **Algorithms:** These bills explore how the algorithms prioritize salacious or harmful content to promote user engagement. Most platforms appear to prefer algorithmic prioritization over other types of categorization, like chronological feeds. Algorithmic curation that promotes—whether intentional or not—harmful content has prompted concern about technology companies' focus on engagement over public welfare. .
- **Common Carriers**: These bills seek to define platforms as common carriers or public utilities, which would require them to treat all legal content equally and refrain from moderation outside of strictly illegal content. These calls have primarily come from Republican politicians, which reflects Republican distrust in platforms' political neutrality. As of 2021, 92% of Republicans believe that social media sites intentionally censor political viewpoints they find objectionable.[19] However, several reputable studies have concluded that right-leaning platforms most frequently publish demonstrably false content.

Questions for the Reader

- Who should determine what constitutes "harm" or "risk" and on what criteria? The government? The platforms? An external third party?
- What online privileges would you be willing to give up to increase the burden of liability on the platforms to curb harmful content and violent organizing? Real-time posting? User-driven moderation? The ability to comment on posts and videos? The ability to share content from sources that haven't been verified by the platforms?
- Would you be willing to give up an algorithm that learns your preferences in exchange for more limited opportunities for those algorithms to show "harmful" content?
- Do you believe platforms are a public right? Should they be treated as public utilities?

- What government entity should be charged with content moderation oversight, and how would protections be put into place to ensure citizens' privacy is safeguarded?
- Can we ensure the standards of what constitutes harm don't change based on which politician or party is in power? How?
- How much should the government attempt to regulate this space? Where is the line between citizen safety and government overreach?

- Should the Republicans succeed in constraining Section 230, how would you determine who is sharing true vs. false information?
- Should the Democrats succeed in constraining Section 230, how would you determine what causes "irreparable harm" to society?

[19] Emily Vogels. "Support for more regulation of tech companies has declined in U.S., especially among Republicans." Pew Research Center, May 13, 2022. https://www.pewresearch.org/fact-tank/2022/05/13/support-for-more-regulation-of-tech-companies-has-declined-in-u-s-especially-among-republicans/.

WHY PROPOSED LEGISLATION IS UNLIKELY TO WORK (OR BE PASSED AT ALL)

America's Blessing and Curse: We Are Bound to the First Amendment

In the United States, in most cases, you can say almost anything about an individual or group of people regarding their gender, race, religion, sexual orientation, or political affiliation. This is the blessing and the curse of the First Amendment.[20] It allows a free flow of ideas and dialogue that question social norms and government power structures—like the #MeToo and #BlackLivesMatter movements—without fear of reprisal from the government.

However, the First Amendment also allows for mis- and disinformation to circulate and take hold in our society. In 2021, the Center for Countering Digital Hate found that just twelve individuals, known as the "Disinformation Dozen," were responsible for 65% of anti-vaccine content online globally.[21]

> If you're upset that Twitter and Facebook keep removing content that favors your political viewpoints,
> Your problem is with the First Amendment, not Section 230.
>
> If you're upset that your favorite social media site won't take down content that offends you,
> Your problem is with the First Amendment, not Section 230.
>
> If you wish social media services had to be politically neutral,
> Your problem is with the First Amendment, not Section 230.
>
> - Jess Miers, TechDirt [22]

[20] James Boyle. "Everything You Know about Section 230 Is Wrong (but Why?)." Techdirt, October 29, 2021. https://www.techdirt.com/2021/10/29/everything-you-know-about-section-230-is-wrong-why/.
[21] Center for Countering Digital Hate. "The Disinformation Dozen: Center for Countering Digital Hate." Accessed May 6, 2022. https://www.counterhate.com/disinformationdozen.
[22] Jess Miers. "Your Problem Is Not with Section 230, but the 1st Amendment." Techdirt, March 9, 2021. https://www.techdirt.com/2020/11/02/your-problem-is-not-with-section-230-1st-amendment/.

Content Moderation at Scale is Difficult to Do Well

> And thus, throwing humility to the wind, I'd like to propose Masnick's Impossibility Theorem, as a sort of play on Arrow's Impossibility Theorem. Content moderation at scale is impossible to do well. More specifically, it will always end up frustrating very large segments of the population and will always fail to accurately represent the "proper" level of moderation of anyone.
>
> - Mike Masnick, TechDirt [23]

Economist Kenneth Arrow's Impossibility Theorem argues there cannot be a perfect voting system that reflects the will of the people. No matter which system is used, there will be an inherent unfairness to it. Editor of TechDirt Mike Masnick believes the same to be true of content moderation.[24]

The vast majority of individuals want moderated digital spaces. No one wants an inbox full of spam or a social platform solely filled with abusive and harassing content. And while we can rely on technology to filter out some words, phrases, and images that directly correlate to the rules, most negative content is harmful within a specific context . Context is extremely difficult to automate and rarely universally agreed upon. So if the experts can't agree, how can we legislate what "over-" or "under-" moderation looks like across a broad variety of platforms?

> "Last year, when we turned an entire conference of 'content moderation' specialists into content moderators for an hour, we found that there were exactly zero cases where we could get all attendees to agree on what should be done in any of the eight cases we presented."
>
> - Mike Masnick, TechDirt[25]

[23] Mike Masnick. "Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well." Techdirt, November 20, 2019. https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/.
[24] Mike Masnick. "Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well." Techdirt, November 20, 2019. https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/.
[25] Ibid.

We Don't Have Enough Data to Objectively Define Digital Social Harms

We know people pay attention to things that are scary and dangerous from an evolutionary perspective, just as they gravitate towards vice from a biological perspective.[26] It's why some algorithms have come under fire—they are built to continually customize their suggestions to the emotions that drive each individual user's behavior online.

In many non-digital cases of scary and dangerous things— such as smoking, drunk driving, and industrial pollution—the government intervened to study their effects and create corresponding bills to address the identified harms. Many advocacy organizations have warned about the "attention economy," in which platforms are incentivized to keep you on their sites or apps as long as possible and by any means possible to optimize their advertising profits. Harmful effects of the attention economy range from increases in teenage depression rates to distrust in our electoral systems based on the rise of conspiracy theories online. However, these claims are often based on external and/or incomplete data. There have not yet been government-mandated, peer-reviewed studies to inform policymakers' ability to define and regulate these psychosocial harms. However, change is on the horizon with legislation proposed in the European Union and in the United States that may compel platforms to make data available for research and oversight purposes.[27]

[26] Nigel Nicholson. "How Hardwired Is Human Behavior?" Harvard Business Review, July 1998. https://hbr.org/1998/07/how-hardwired-is-human-behavior.

[27] Brandie Nonnecke and Camille Carlton, "EU and US legislation seek to open up digital platform data," Science, Feb. 11, 2022, 375(6581): 610-612, https://www.science.org/doi/10.1126/science.abl8537.

FOCUSING ON THE FUTURE: BEWARE OF SECOND-ORDER EFFECTS

Some of the proposed legislation outlined above could fundamentally change the Internet and how we use it to find information and interact with others. On top of incomplete data, much of the proposed legislation appears to be based on simplistic cause and effect analysis that fails to take damaging second- and third-order effects into account. For example, some argue that revoking Section 230 protections will force platforms to moderate every post, video, or image, ridding the Internet of "bad" content. However, regulation is never that simple; it comes with trade-offs. In that hypothetical case, forcing platforms to accept liability for third-party content would probably prompt some to stop allowing third-party content altogether given the high stakes and the costs of moderation at that scale.

Scenarios Analysis

We conducted a scenarios analysis exercise to think through what the Internet might look like if two of the key points of contention in the content moderation debate change: transparency and the level and scope of moderation. Government-mandated adjustments to these factors have the potential to drive the most dramatic evolution of the content moderation landscape.

This exercise is intended to improve understanding of what an online world is likely to look if transparency and moderation tactics change. This exercise is not intended to predict the individual effects of legislative or regulatory change; rather, it is an effort to explore the broader implications of these tactics. This allows us to focus on how to navigate the uncertainties of the highly dynamic tech policy space without focusing on specific details that may or may not play out as we have described.

Our scenarios analysis considers six distinct factors to help provide a well-rounded view of each potential future: political, economic, social, security, technological, and environmental/health. The center of the four quadrants is our origin point—how each of us uniquely views the online content ecosystem in 2022 given there are no universal definitions of moderation and transparency as they relate to the Internet. One person may see a platform like Twitter as highly moderated, while another sees it as barely moderated.

For the purposes of this exercise, new government policies would have the following effects on moderation and transparency:

- **High Moderation:** Platforms more actively monitor, flag, and remove content than the current status quo.
- **Low Moderation:** Platforms less actively monitor, flag, and remove content than the current status quo.
- **High Transparency:** Platforms increase self-reporting and external access to platform information than the current status quo.
- **Low Transparency:** Platforms decrease self-reporting and external access to platform information than the current status quo.

SCENARIO NARRATIVES

High Moderation

Scenario 1: Platforms as Judge
No changes to Section 230 have passed, but platforms are responding to increased public pressure to moderate aggressively. Differences between platform options have become more pronounced as they increase both their technological and human capacity for moderation. However, this increased scrutiny has led to platforms pulling back on the public reporting and research partnerships due to a lack of incentive. The claim that platforms are swaying elections increases based on that lack of transparency, which has deepened partisan distrust in the US electoral system.

Scenario 2: Government Watchdogs
The US Government has required platforms to tightly moderate many different categories of illegal and unprotected speech, image, and video content. Platforms are also legally obligated to be fully transparent about why and how they make their moderation decisions. Given the range of regulated content, platforms no longer allow real-time posting; each new comment, image, and/or video is subject to review. The US tech economy becomes less profitable and innovative, creating room for Chinese companies to dominate the global economy.

Low Transparency — **Status Quo** — **High Transparency**

Scenario 3: More Money, More Politics
Congress has backed off its regulatory push against Section 230 and turned their focus toward antitrust with the idea that the fear of being broken up will incentivize good behavior. Platforms have begun to pour even more money into politics. They develop increasingly close ties with the police to show the benefits of the large scale data they provide, provoking a mass outcry about the human rights implications of sharing user data. Users are disillusioned with both policymakers and platforms.

Scenario 4: User-Driven and Distributed
Congress has passed legislation that requires full and detailed transparency about what platforms allow online. The federal government itself will not prosecute, but has opened the door to civil liability should users feel they have been wronged by moderation. Given the requirement for clear and consistent rules, many platforms default to under-moderating (other than clearly illegal content) in favor of free speech to minimize potential for liability.

Low Moderation

Contemplating Trade-Offs

As you consider each of the four scenarios, challenge yourself to think through each in the role of a digital citizen, government official, and in-house content moderator. In taking on each role, consider the different opportunities and constraints you face and how these affect your actions.

Questions for the Reader

- What future scenario leads to your ideal vision for the web?
- What trade-offs would you be willing to make politically, socially, economically and technologically to get there?
- What happens in the best case for each scenario?
- Who are the winners and the losers in each scenario? Does it feel fair and equitable to you?

Scenario 1: Platforms as Judge (high moderation / low transparency)
No changes to Section 230 have passed, but platforms are responding to increased public pressure to moderate aggressively. Differences between platform options have become more pronounced as they increase both their technological and human capacity for moderation. However, this increased scrutiny has led to platforms pulling back on the public reporting and research partnerships due to a lack of incentive. The claim that platforms are swaying elections increases based on that lack of transparency, which has deepened partisan distrust in the US electoral system.

- Large platforms hire more moderators with language, geographic, and issue area expertise to better identify harmful content based on cultural and social contexts.
- Large companies also improve investment into AI capabilities and the application of automated decision making.
- The ability to choose specific sites and apps creates more echo chambers, increasing polarization. Misinformation, radical rhetoric, and other harmful content proliferates on these platforms before the government shuts them down due to moderation violations, creating a whack-a-mole effect.
- Many platforms conduct backroom deals with authoritarian or otherwise problematic governments. Some internal company cultures keep platforms from building these relationships, although it is often difficult for the government or public to determine which companies are engaging in such behaviors.

Scenario 2: Government Watchdogs (high moderation / high transparency)
The US Government has required platforms to moderate many different categories of speech, image, and video content. Platforms are also legally obligated to be fully transparent about why and how they make their moderation decisions. Given the range of regulated content, platforms no longer allow content to be posted in real-time; each new post, comment, image, and/or video is subject to review. The US tech economy becomes less profitable and innovative, creating room for Chinese companies to dominate the global economy.

- A new federal agency is created and generations of government bureaucrats become specialists in the many distinct categories of content.
- Platforms with sufficient resources have invested in AI and are able to speed up the review process. Less profitable platforms are forced out of the market due to slow review times, making it difficult for new companies to enter without significant venture capital investment.
- Small platforms routinely appear to subvert regulation and engage in prohibited forms of discussion or content creation. These platforms often disappear without notice, but do outsized harm while active.
- News commentators, both in the US and abroad, opine on whether regulation has drifted into censorship and if the US is backsliding on its democratic values. Others praise the system for ridding the Internet of harmful and dangerous content.

Scenario 3: More Money, More Politics (low moderation / low transparency)
Congress has backed off its regulatory push against Section 230 and turned their focus toward antitrust with the idea that the fear of being broken up will incentivize good behavior. Platforms, in response, have begun to pour even more money into lobbying, and they develop increasingly close ties with law enforcement and intelligence agencies to show the benefits of the large-scale data they provide, provoking a mass outcry about the human rights implications of sharing user data. Users are disillusioned with both policymakers and the platforms.

- There is a large uptick in anecdotal evidence of body image concerns, racial and political divides, and suicides because of online harassment, although researchers lack data to back up claims given the lack of transparency.
- US allies publicly express concern about the lack of US reform and increase their own regulation, ban US sites, and/or create walled gardens to keep their citizens safe from the uptick in harmful content.
- Countries that are antagonistic to the US point to it as an example of moral corruption.
- Platforms turn their focus toward web3 and the metaverse, with the hope that newer versions of their services will result in a different relationship with users and policymakers.

Scenario 4: User-Driven and Distributed (low moderation / high transparency)
Congress has passed legislation that requires full and detailed transparency about what platforms allow online. The federal government itself will not prosecute but has opened the door to civil liability should users feel they have been wronged by moderation. Given the requirement for clear and consistent rules, many platforms default to under-moderating (other than clearly illegal content) in favor of free speech to minimize potential for liability.

- Academics, journalists, and policymakers gain unprecedented levels of access to platform data resulting in a greater understanding of the trends and impacts of online activities.
- An initial flurry of lawsuits results in some consistent precedents set, but even platforms marketing themselves as "safe online spaces" hoping to appeal to people looking for a more positive user experience face consistent challenges with community norms and lawsuits impeding their growth.
- Negative content like hate speech and misinformation proliferate in the developing world because there's less pressure for platforms to acquire foreign language capabilities. Hate crimes and political conflict, similar to the genocide in Myanmar,[28] spike.
- The difference in dynamics and resources between richer and poorer countries creates a digital divide with stark differences in user experiences.

[28] The Guardian. "Rohingya sue Facebook for £150bn over Myanmar genocide." Accessed May 23, 2022. https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

Reflecting on the Unknown

Each of the potential futures described above comes with trade-offs, uncertainties, and unintended consequences we can't yet fully grasp. The scenarios were built with our understanding of the Internet in 2022; new technologies, changes in US political dynamics, and foreign Internet regulation are just a few dynamics that are likely to shift the scenarios we've outlined. Please consider the following questions as you reflect on your version of better platform transparency and content moderation based on the above scenarios.

Questions for the Reader

- What role will platforms play in society 10 years from now? Twenty years from now?
- What new types of harmful content might these platforms host?
- How will the evolution of our most commonly used products—smartphones, tablets, smartwatches, computers, etc.—change how people interact with content?
- How will current companies adapt to changes in regulation and technology?
- How will new competitors contribute to or lessen our content moderation dilemmas?
- How will geopolitical and market dynamics affect these futures?
- Will this regulation constrain or foster innovation?
- How will younger generations feel about these challenges and approach regulation?

Platforms have had a society-shaping influence of a magnitude inconceivable only a few decades ago, and in the scenarios above, we see the potential for legislation to dramatically change our lives across political, economic, social, security, technological, and environmental/health dimensions. However, change is inevitable. As we work toward responsible and effective legislation, we offer a series of recommendations to continue progress toward a better platform transparency and content moderation.

RECOMMENDATIONS

Recommendation 1: Maximize User Control

> “It is the policy of the United States…to encourage the development of technologies which maximize user control over what information is received by individuals…who use the internet.”
>
> - From Section 230 of the Communications Decency Act[29]

If the First Amendment encourages freedom of expression, does Section 230 encourage freedom of impression? In their piece *Delegability, Or, The Twenty-Nine Words that the Internet Forgot*, Richard Reisman and Chris Riley interpret the text of Section 230 as “an aspirational statement of individual agency over the receipt, not the production, of information.” [30] This idea is best illustrated through three primary concepts:

- **Portability** gives the user the ability to securely transfer their own data to another user or a competing platform, breaking the monopolistic cycle of network effects. This means smaller platforms could have a chance to compete against the incumbent players as users move their data from one platform to another.
- **Interoperability** gives the user the ability to seamlessly communicate with a user of a competing platform, similar to the ability to call anyone regardless of the phone plan they use.
- **Delegatability** gives the user the ability to delegate authority to “infomediaries” (*info*rmation inter*mediary)*—or companies that serve as a middleman between you and the platform. Users would be able to choose the filters and experience they could have on the Internet, while these companies handle the technical integration and compliance with the platforms.[31]

These concepts first gained traction in 2019 with the introduction of the bi-partisan ACCESS Act, which was recently reintroduced in 2022 with the goal of putting “consumers in the driver’s seat when it comes to how and where they use social.”[32] The intention is to move beyond the debates surrounding the legality of the content on the Internet and focus more on giving consumers greater choice and access to their own experience.

[29] “47 U.S. Code § 230 - Protection for Private Blocking and Screening of Offensive Material.” Legal Information Institute. Accessed May 5, 2022. https://www.law.cornell.edu/uscode/text/47/230.
[30] Richard Reisman and Chris Riley. “Delegation, or, the Twenty Nine Words That the Internet Forgot.” R Street, March 4, 2022. https://www.rstreet.org/2022/02/28/delegation-or-the-twenty-nine-words-that-the-internet-forgot/.
[31] Richard Reisman and Chris Riley. “Delegation, or, the Twenty Nine Words That the Internet Forgot.” R Street, March 4, 2022. https://www.rstreet.org/2022/02/28/delegation-or-the-twenty-nine-words-that-the-internet-forgot/.
[32] Mark R. Warner. “The Access Act of 2022.” Accessed on June 4, 2022. https://www.warner.senate.gov/public/_cache/files/9/f/9f5af2f7-de62-4c05-b1dd-82d5618fb843/BA9F3B16A519F296CAEDE9B7EFAB0B7A.access-act-one-pager.pdf.

The idea of a customizable web experience lends itself to the concern that filter bubbles will worsen. The counterargument is that this filtering system would actually serve to sideline those trolls into smaller communities that wouldn't have the opportunity to attract new and vulnerable users. While we don't have to like it, we do have to acknowledge that those with divisive viewpoints have the right to express those viewpoints.[33]

Recommendation 2: Use Data to Inform Research and Oversight

> "Clearly, we can't rely on the platforms alone to reduce these harms on their own. Members of this committee will understand that we need to base solutions on independent, rigorous, scientific inquiry based on concrete data…Every day that my team cannot access the data needed to do their research puts us further behind in the race to find answers…It's time for Congress to act to ensure that researchers, journalists, and the public have access to the data we need to both study online misinformation and build real solutions."
>
> - Laura Edelson, Congressional Testimony, September 2021[34]

The biggest names in tech have used targeted advertising to fund their platforms and generate revenue, incentivizing them to create and leverage an immense amount of user data. Google, for example, launched AdWords in 2000, and by 2021, it was a $209.5B business.[35] Tracking users' activity to create search suggestions and targeted ads, coupled with the default settings that allow Google to access users' locations, has allowed Google to build shockingly accurate profiles. A journalist in early 2022 downloaded all the data Google had accumulated on her, which amounted to 2 GB or roughly 1.5 million word documents.[36]

Improving researcher access to the wealth of data tech companies have accumulated for empirical research will allow policymakers to produce evidence-based proposals.[37] There is currently no government mandate for the platforms to enter into research partnerships or a way to hold platforms accountable when data reveal wrongdoings.

[33] Mike Masnick. "Protocols, Not Platforms: A Technological Approach to Free Speech." Knight First Amendment Institute, August 21, 2019. https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

[34] U.S. House of Representatives Committee on Science, Space and Technology. "Testimony of Laura Edelson, NYU Cybersecurity for Democracy." September 28, 21, https://science.house.gov/imo/media/doc/Edelson%20Testimony.pdf.

[35] Joseph Johnson. "Google: AD Revenue 2001-2020." Statista, February 7, 2022. https://www.statista.com/statistics/266249/advertising-revenue-of-google/#:~:text=In%202021%2C%20Google's%20ad%20revenue,and%20apps

[36] Nicole Martin. "How Much Does Google Really Know about You? A Lot." Forbes Magazine, March 11, 2019. https://www.forbes.com/sites/nicolemartin1/2019/03/11/how-much-does-google-really-know-about-you-a-lot/?sh=66678e247f5d.

[37] Brandie Nonnecke and Camille Carlton, "EU and US legislation seek to open up digital platform data," Science, Feb. 11, 2022, 375(6581): 610-612, https://www.science.org/doi/10.1126/science.abl8537.

Two bills that attempt to address the data and research gap are compelling:

- The **Platform Accountability and Transparency Act (PATA)**, proposed in December 2021, addresses the issue of researcher access. The bill mandates that platforms make data available to qualified researchers who are affiliated with a university and pursuing projects that have been approved by the National Science Foundation (NSF). The government would establish the Platform Accountability and Transparency Office (PATO—responsible for the security of the data access and transfer between the researchers and platform—under the Federal Trade Commission (FTC) .[38]
- The **Algorithmic Accountability Act of 2022** calls for impact assessments in which platforms would disclose the algorithm training dataset and demographics, expected and measured ethical and social consequences, and steps taken to consult with impacted communities. These assessments would inform an annual FTC anonymized trend report. It would also establish a repository of information where consumers and advocates can review which critical decisions have been automated by companies tracked to data sources, high level metrics, and how to contest decisions.[39]

Similar to food safety, controlled substances, waste management and environmental harms, these types of data collection and impact assessments would hold companies accountable for social harms. The difference is that with those examples, it is possible to conduct independent studies that establish clear and factual connections to health challenges. We don't have a fully representative and universally accepted view of how social media platforms amplify content and tailor suggestions for users because of privacy and First Amendment protections. Instead, we can focus on understanding how platform algorithms influence what we buy, what we believe, and what we engage with.[40] With that knowledge, policymakers can create effective legislation and users can apply pressure on the companies to create safer platforms.

Recommendation 3: Focus on "Friction"

Even if we're able to collect meaningful data, craft tailored and influential legislation, and curb the amount of harmful content, people will continue to engage in bad behavior online. Some of this behavior is clearly intentional; bad actors routinely upload child sexual abuse material, spread hateful rhetoric, and engage in cyberbullying. But there are also plenty of people who don't realize that what they're doing contributes to a negative digital culture and environment. Some who spread misinformation may truly believe in what they're sharing or not recognize it as false.

[38] Congress.gov. "S.797 - PACT Act." March 17, 202. https://www.congress.gov/bill/117th-congress/senate-bill/797/text; Brandie Nonnecke and Camille Carlton, "EU and US legislation seek to open up digital platform data," Science, Feb. 11, 2022, 375(6581): 610-612, https://www.science.org/doi/10.1126/science.abl8537.

[39] Ron Wyden. "[Summary] Algorithmic Accountability Act of 2022." February 3, 2022. https://www.wyden.senate.gov/imo/media/doc/2022-02-03%20Algorithmic%20Accountability%20Act%20of%202022%20One-pager.pdf.

[40] Jacob Metcalf, Brittany Smith, and Emanuel Moss. "A New Proposed Law Could Actually Hold Big Tech Accountable for Its Algorithms." Slate Magazine, February 9, 2022. https://slate.com/technology/2022/02/algorithmic-accountability-act-wyden.html.

While we work on creating better solutions for platforms, civil society, and governments, how can we encourage people to be better digital citizens to help mitigate bad behavior? What would incentivize people to care more about their online spaces?

In James Madison's Federalist papers, he states that "where no substantial occasion presents itself, the most frivolous and fanciful distinctions have been sufficient to kindle their unfriendly passions and excite their most violent conflicts."[41] The goal in drafting the US Constitution was to create friction points to ensure that the core elements of a sustainable democracy were not overly malleable to changing political passions.

According to social psychologist Jonathan Haidt, there are three major forces that create successful democracies:

- Social capital (extensive social networks with high levels of trust)
- Shared stories
- Strong institutions

In his piece, *Why the Last Ten Years of American Life Have Been Uniquely Stupid*, Haidt explores how all three have been systematically eroded by social media's focus on virality rather than the friction Madison knew was necessary for a civil functional society.[42]

With the introduction of shares, likes, and retweets, social media platforms moved away from the goal of strengthening these necessary social ties and towards the goal of increased engagement by the masses. Algorithms were developed to deliver users content they predicted would generate a "like" or interaction. The problem was that when users began to understand that the more triggering and divisive content got the most attention, they began to post accordingly. The commonly used term "engagement equals enragement" sums up the fact that social media has been weaponized to trigger our emotions, against each other, our institutions, and our democracy.

In her testimony to Congress, Facebook whistleblower Frances Haugen pointed to simple "viewpoint neutral" changes that could influence the underlying nature of how we interact with each other online without infringing on First Amendment rights. These include modifying the "share" function on Facebook so that after any content has been shared twice, the third person in the chain must take the time to copy and paste the content into a new post in order to slow the spread of content.[43]

[41] James Madison. "Federalist Papers No. 10 (1787)." Bill of Rights Institute. Accessed May 6, 2022. https://billofrightsinstitute.org/primary-sources/federalist-no-10.

[42] Jonathan Haidt. "Why the Past 10 Years of American Life Have Been Uniquely Stupid." The Atlantic, April 20, 2022. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/.

[43] Frances Haugen. "Statement of Frances Haugen." Whistleblower Aid, October 4, 2021. "https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49.

The Social Media NUDGE Act is a 2022 bi-partisan bill co-sponsored by Sens. Amy Klobuchar (D-MN) and Cynthia Lummis (R-WY). The Act empowers the NSF and the National Academies of Sciences, Engineering, and Medicine to develop "content neutral" approaches to adding friction to social media platforms. The FTC would oversee the enforcement of those approaches, and platforms that violate the rules would be subject to prosecution based on unfair and deceptive practices.[44]

Encouraging experimentation with clear transparency about how friction is implemented and how it affects users will help ensure more equitable outcomes. The shadow side of content neutrality is that while these measures may break the virality of misinformation, they may also break the virality of important social movements like #MeToo. We must begin to create an open dialogue between the government, users, and platforms to determine the right way to hold and pass the microphone to keep society growing and thriving.

Recommendation 4: Educate the Next Generation

A 2019 Stanford University study found that two-thirds of high school students could not tell the difference between a news story and a sponsored ad, and 52% believed that a low-quality video claiming to show ballot stuffing in the US (actually shot in Russia) was "strong evidence" of voter fraud, despite an abundance of easily discoverable articles that negated the claim.[45] These results are troubling and underscore the urgent need for high-quality digital literacy education.

Thirteen states and counting have introduced or passed legislation in 2022 focused on improving digital literacy and digital citizenship.[46] In 2019, Senator Amy Klobuchar (D-MN) proposed the Revive Digital Citizenship and Media Literacy Act, which would create a grant program to develop state-wide media literacy education guidelines, incorporate it into curriculums, hire experienced media literacy educators, and promote media literacy training for educators.[47] Reviving this bill or drafting a similar piece of legislation would establish a common baseline for online content education and arm the next generation with better tools to combat mis-/disinformation and other online social harms. Creating an informed public capable of community, civic, and democratic engagement starts with educating our youth to take responsibility for the content they create, share, and consume.

[44] Issie Lapowsky. "New Bill Would Force Social Media Giants to Embrace Friction - or Else." Protocol, February 11, 2022. https://www.protocol.com/bulletins/social-media-nudge-act.

[45] Joel Breakstone, Mark Smith and Sam Wineburg. "Civic Online Reasoning National Portrait - Stanford University." Stanford History Education Group. Accessed May 6, 2022. https://stacks.stanford.edu/file/druid:gf151tb4868/Civic%20Online%20Reasoning%20National%20Portrait.pdf.

[46] EveryLibrary. "Good Bills We're Following this Session." April 2, 2022. https://www.everylibrary.org/good_bills_this_session.

[47] Congress.gov. "S.2240 - 116th Congress (2019-2020): Digital Citizenship and Media Literacy Act." Accessed May 6, 2022. https://www.congress.gov/bill/116th-congress/senate-bill/2240.

CONCLUSION

Section 230 and the general content moderation landscape are complex topics, even for the most well-versed scholars and policy experts. Our society lacks a shared understanding of what constitutes harmful content, in addition to the wide variety of (and sometimes uninformed) perspectives about how to approach the problem. Given the high social, political, economic, and health stakes of regulation, it is not a decision we should make quickly or lightly. Creating mechanisms that build trust between and educate all stakeholders is a necessary component of legislative reform. And imparting a certain amount of responsibility onto users will alleviate some of the burden on civil society, government, and platforms. It's crucial that we all play our part in supporting more responsible platforms.

ABOUT

This project was supported by the CITRIS Policy Lab and Our Better Web.

The CITRIS Policy Lab is a sub-organization of the Center for Information Technology Research in the Interest of Society and the Banatao Institute (CITRIS), headquartered on the UC Berkeley campus. The CITRIS Policy Lab supports interdisciplinary research regarding the role of formal and informal regulation in promoting innovation and amplifying positive effects.

Our Better Web is an interdisciplinary initiative at UC Berkeley that seeks to strengthen the power of the Internet to support resilient, democratic societies and address the sharp rise of online harms.

The views, findings, and suggestions in this report are based on research conducted through the CITRIS Policy Lab and are not affiliated with the authors' employers or any other organization.

AUTHORS

Amy Benziger is a graduate of the UC Berkeley Master of Public Affairs program. She is a Vice President at Breakwater Strategy, a strategic communications and insights agency that advises companies, brands, coalitions, and non-profits on how to navigate change, crises, and complexities. Prior to that, she was Head of Projects at Cisco's Innovation Lab (CHILL).

Chelsea Magnant is a graduate of the UC Berkeley Master of Public Affairs program. She is a Manager of Government Affairs and Public Policy at Google. Before Google, she spent nine years at the Central Intelligence Agency as a political analyst.

ADVISORS

Hany Farid is a professor in Electrical Engineering & Computer Sciences and the School of Information at UC Berkeley. He is a leading expert in image analysis, digital forensics, forensic science, misinformation, and the intersection of technology and society particularly as it pertains to online harms.

Brandie Nonnecke, PhD, is Founding Director of UC Berkeley's CITRIS Policy Lab where she supports interdisciplinary tech policy research and engagement. She is a Technology and Human Rights Fellow at the Carr Center for Human Rights Policy at the Harvard Kennedy School and a Fellow at the Schmidt Futures International Strategy Forum.

Special thanks to Alan Kyle, UC Berkeley Master of Information Management Graduate Student, for database research.

CREDIT
Emoji icons created by Freepik - Flaticon. Cover image from Unsplash.com.