Texas A&M Law Review

Volume 8 | Issue 3

4-30-2021

Regulatory Goldilocks: Finding the Just and Right Fit for Content Moderation on Social Platforms

Nina Brown
Syracuse University, nmibrown@syr.edu

ARTICLES

REGULATORY GOLDILOCKS: FINDING THE JUST AND RIGHT FIT FOR CONTENT MODERATION ON SOCIAL PLATFORMS

*by: Nina I. Brown**

ABSTRACT

Social media is a valuable tool that has allowed its users to connect and share ideas in unprecedented ways. But this ease of communication has also opened the door for rampant abuse. Indeed, social networks have become breeding grounds for hate speech, misinformation, terrorist activities, and other harmful content. The COVID-19 pandemic, growing civil unrest, and the polarization of American politics have exacerbated the toxicity in recent months and years.

Although social platforms engage in content moderation, the criteria for determining what constitutes harmful content is unclear to both their users and employees tasked with removing it. This lack of transparency has afforded social platforms the flexibility of removing content as it suits them: in the way that best maximizes their profits. But it has also inspired little confidence in social platforms' ability to solve the problem independently and has left legislators, legal scholars, and the general public calling for a more aggressive—and often a government-led—approach to content moderation.

The thorn in any effort to regulate content on social platforms is, of course, the First Amendment. With this in mind, a variety of different options have been suggested to ameliorate harmful content without running afoul of the Constitution. Many legislators have suggested amending or altogether repealing section 230 of the Communications Decency Act. Section 230 is a valuable legal shield that immunizes internet service providers—like social platforms—from liability for the content that users post. This approach would likely reduce the volume of online abuses, but it would also have the practical effect of stifling harmless—and even socially beneficial—dialogue on social media.

While there is a clear need for some level of content regulation for social platforms, the risks of government regulation are too great. Yet the current self-regulatory scheme has failed in that it continues to enable an abundance of harmful speech to persist online. This Article explores these models of regulation and suggests a third model: industry self-regulation. Although there is some legal scholarship on social media content moderation, none explore such a model. As this Article will demonstrate, an industry-wide governance model is the optimal solution to reduce harmful speech without hindering the free exchange of ideas on social media.

DOI: https://doi.org/10.37419/LR.V8.I3.1

* Assistant Professor, S.I. Newhouse School of Public Communications; J.D., Cornell Law School. Many thanks to Amanda Nardozza for her assistance in the preparation of this work. Thank you also to the editors of the *Texas A&M Law Review* for excellent editorial assistance.

TABLE OF CONTENTS

I. INTRODUCTION

Not long after the COVID-19 pandemic reached the United States, false information about the virus began appearing on social platforms. Videos and articles spread across Facebook and Twitter that touted at-home "breath-holding" tests as reliable alternatives to medical testing.[1] A false memo circulated on Facebook claimed that consuming alcoholic drinks prevented the coronavirus.[2] False claims suggesting the virus was caused by and transmitted through 5G networks sur-

1. Audrey Cher, *Don't Hold Your Breath. Experts Debunk Dangerous Myths About the Coronavirus*, CNBC (May 1, 2020, 3:29 AM), https://www.cnbc.com/2020/05/01/experts-explain-why-coronavirus-myths-misinformation-can-be- dangerous.html [https://perma.cc/N4M2-TFQR].

2. Arijeta Lajka, *Consuming Alcoholic Drinks Does Not Prevent the Coronavirus*, AP NEWS (Mar. 31, 2020), https://apnews.com/afs:Content:8709292273 [https://perma.cc/Q3AK-UR2E].

faced across social platforms.[3] That these falsehoods could spread so easily is no surprise: social media has long been a breeding ground for misinformation and other harmful speech.[4] The surprise was that social platforms promised to take a proactive role in removing false and potentially harmful information related to the coronavirus.

Facebook was first. It announced it would warn users after they interacted with posts containing "harmful" coronavirus misinformation[5] and link those users to resources from the World Health Organization, the Centers for Disease Control, and local health authorities to combat the false information.[6] Other platforms followed suit. YouTube removed thousands of videos containing false information about the coronavirus.[7] Twitter began labeling false or misleading tweets about the coronavirus.[8]

This moderation marked a departure from the laissez-faire approach social platforms have traditionally taken regarding false and harmful content. The lack of regulation for social platforms has allowed them to prefer models that prioritize the "free speech" of their users despite longstanding cries for safer online spaces.[9] With few ex-

3. Matthew Brown, *Fact Check: 5G Technology Is Not Linked to Coronavirus*, USA TODAY (Apr. 29, 2020, 2:52 PM), https://www.usatoday.com/story/news/factcheck/2020/04/23/fact-check-5-g-technology-not-linked-coronavirus/3006152001/ [https://perma.cc/BM4M-N8K9].

4. Nina I. Brown & Jonathan Peters, *Say This, Not That: Government Regulation and Control of Social Media*, 68 SYRACUSE L. REV. 521, 525 (2018).

5. Shannon Bond, *Did You Fall for a Coronavirus Hoax? Facebook Will Let You Know*, NPR (Apr. 16, 2020, 9:00 AM), https://www.npr.org/2020/04/16/835579533/did-you-fall-for-a-coronavirus-hoax-facebook-will-let-you-know [https://perma.cc/9LAY-T5YP]; *Keeping People Informed, Safe, and Supported on Instagram*, INSTAGRAM: OFF. BLOG (Mar. 24, 2020), https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram [https://perma.cc/8QJB-R9R3].

6. Kang-Xing Jin, *Keeping People Safe and Informed About the Coronavirus*, FACEBOOK: NEWSROOM (June 24, 2020, 5:00 AM), https://about.fb.com/news/2020/05/coronavirus/ [https://perma.cc/J26M-KJD4].

7. Ina Fried, *YouTube Pulls Coronavirus Misinformation Videos*, AXIOS (Apr. 7, 2020), https://www.axios.com/youtube-coronavirus-misinformation-videos-google-d9ce89cb-0de0-4f50-8a25-5923e078a858.html [https://perma.cc/P6A7-NB63].

8. Rebecca Heilweil, *Twitter Now Labels Misleading Coronavirus Tweets with a Misleading Label*, VOX (May 11, 2020, 3:55 PM), https://www.vox.com/recode/2020/5/11/21254889/twitter-coronavirus-covid-misinformation-warnings-labels [https://perma.cc/759M-N7DJ]; Yoel Roth & Nick Pickles, *Updating Our Approach to Misleading Information*, TWITTER BLOG (May 11, 2020), https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html [https://perma.cc/PS3E-RQWF].

9. *See* Richard L. Hasen, *Cheap Speech and What It Has Done (to American Democracy)*, 16 FIRST AMEND. L. REV. 200, 226 (2017) (quoting ZEYNEP TUFECKI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST 267 (2017)); *see also* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1618 (2018) (discussing how social platforms' content-moderation policies have been influenced by the concerns about user free speech and collateral censorship); Michael Holmes, *ISIS Looking for Recruits Online*, WWLP (June 21, 2014, 12:29 AM), http://wwlp.com/2014/06/20/isis-

ceptions, social platforms are under no legal obligation to police or remove harmful content. [10] Additionally, a powerful federal law—section 230 of the Communications Decency Act ("CDA")—immunizes social platforms for harms resulting from user-generated content.[11] This combination of immunity and lack of regulatory oversight is what has enabled social platforms to function: their business model depends on users freely creating and uploading content—driving advertising revenue—with little risk of liability for the publisher of that content—the platform.[12]

The practical result is that any content regulation that exists on social platforms is driven by the platform itself. Typically, this is driven by the platform's terms of use, which often prohibit posting and sharing certain harmful material. However, because platforms are under no legal obligation to remove posts that violate the terms, they are free to enforce—or ignore—abuses of the terms at will. As a result, this self-regulatory framework has allowed social platforms to develop policies that prioritize profits over safety.[13]

These policies have had serious consequences. Facebook's lenient stance on hate speech has helped fuel the proliferation of white supremacists and other extremist groups and actors, at times resulting in tangible physical harm.[14] YouTube's and Twitter's policies have al-

looking-for-recruits-online/ [http://perma.cc/2E4Y-25PB] (noting that Twitter founder Biz Stone—who is no longer with the company—responded to media questions about ISIS's use of Twitter to publicize its acts of terrorism by saying, "[i]f you want to create a platform that allows for the freedom of expression for hundreds of millions of people around the world, you really have to take the good with the bad"); Somini Sengupta, *Twitter's Free Speech Defender*, N.Y. TIMES (Sept. 2, 2012), http://www.nytimes.com/2012/09/03/technology/twitter-chief-lawyer-alexander-macgillivray-defender-free-speech.html [https://perma.cc/X39J-LZGJ]; Mike Isaac, Cecilia Kang & Sheera Frenkel, *Zuckerberg Defends Hands-Off Approach to Trump's Posts*, N.Y. TIMES (June 3, 2020), https://www.nytimes.com/2020/06/02/technology/zuckerberg-defends-facebook-trump-posts.html [https://perma.cc/SJU2-VNEZ].

10. A notable exception is the Digital Millennium Copyright Act, which has a notice and takedown provision that requires a service provider to remove infringing material once the provider is on notice of its existence to merit safe harbor from copyright infringement charges. 17 U.S.C. § 512(g)(1)–(4).

11. 47 U.S.C. § 230(c).

12. Casey Newton, *Everything You Need to Know About Section 230: The Most Important Law for Online Speech*, THE VERGE (Dec. 29, 2020, 4:50 PM), https://www.theverge.com/21273768/section-230-explained-internet-speech-law-definition-guide-free-moderation [https://perma.cc/9SB9-N6ZH]; *see also* Michael Patty, *Social Media and Censorship: Rethinking State Action Once Again*, 40 MITCHELL HAMLINE L.J. PUB. POL'Y & PRAC. 99, 134 (2019).

13. Chloé Nurik, *"Men Are Scum": Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook*, 13 INT'L J. COMMC'N, 2878, 2880, 2887, 2889 (2019).

14. Daniela Hernandez & Parmy Olson, *Isolation and Social Media Combine to Radicalize Violent Offenders: Social Media Is Increasingly Playing a Role, Especially Among Lone Actors Like the Ones Responsible for El Paso and Dayton Shootings*, WALL ST. J. (Aug. 5, 2019, 5:44 PM), https://www.wsj.com/articles/isolation-and-social-media-combine-to-radicalize-violent-offenders-11565041473 [https://perma.cc/47YE-MKJ3]; Jenni Marsh & Tara Mulholland, *How the Christchurch Terrorist Attack*

lowed terrorists to use their platforms to recruit, spread propaganda, and raise funds.[15] Revenge pornography has spread across Instagram, Facebook, and Twitter.[16] Social platforms were famously exploited by disinformation campaigns during the 2016 U.S. Presidential election, and commentators roundly agree that social platforms did not do enough to reduce, let alone eliminate, disinformation in time for the 2020 presidential election.[17]

Despite the persistent public outcry for a more aggressive response to harmful speech, social platforms have largely abdicated this responsibility.[18] This is not to suggest that they have ignored the problem: each major platform has committed resources to reducing harmful speech and has made progress.[19] However, these shifts have largely been reactive: social platforms have not embraced the concept of proactively reducing abusive content.[20] Only after the most egregious abuses—and particularly following threatened legal action or loss of advertisers—have social platforms responded by removing content, making (often minor) policy changes, deleting user accounts, or amending terms of service.[21]

Was Made for Social Media, CNN BUS., https://www.cnn.com/2019/03/15/tech/christchurch-internet-radicalization-intl/index.html (Mar. 16, 2019, 5:30 PM) [https://perma.cc/F8Z5-CCKP].

15. Crosby v. Twitter, Inc., 921 F.3d 617, 620 (6th Cir. 2019); Nina I. Brown, *Fight Terror, Not Twitter: Insulating Social Media from Material Support Claims*, 37 LOY. L.A. ENT. L. REV. 1, 2 (2016).

16. *See, e.g.*, GoDaddy.com, LLC v. Toups, 429 S.W.3d 752, 753 (Tex. App.—Beaumont 2014, pet. denied) (dismissing claims involving revenge pornography against GoDaddy under section 230); Caitlin Kelly, *Facebook's Anti-Revenge Porn Tools Failed to Protect Katie Hill*, WIRED (Nov. 18, 2019, 11:30 AM), https://www.wired.com/story/katie-hill-revenge-porn-facebook/ [https://perma.cc/W9SY-UCFV] (reporting that intimate photos of former U.S. Representative Katie Hill and her former partner were disseminated across Facebook and Twitter); Olivia Solon, *Inside Facebook's Efforts to Stop Revenge Porn Before It Spreads*, NBC NEWS (Nov. 19, 2019, 10:15 AM), https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631 [https://perma.cc/GY35-JRT3] (reporting that revenge porn targeting twenty-two-year-old Michaela Zehara spread across Instagram in 2016).

17. *See* Brian Beyersdorf, *Regulating the "Most Accessible Marketplace of Ideas in History": Disclosure Requirements in Online Political Advertisements After the 2016 Election*, 107 CALIF. L. REV. 1061, 1082, 1098 (2019).

18. *See* Casey Newton, *YouTube Expands Anti-Harassment Policy to Include All Creators and Public Figures*, THE VERGE (Dec. 11, 2019, 9:00 AM), https://www.theverge.com/2019/12/11/21005185/youtube-harassment-policy-update-malicious-expression-public-figures-maza-crowder [https://perma.cc/SE85-TNZK].

19. *See id.*; Salvador Rodriguez, *Zuckerberg: Facebook Will Prohibit Hate Speech in Its Ads*, CNBC (June 26, 2020, 5:20 PM), https://www.cnbc.com/2020/06/26/zuckerberg-facebook-will-prohibit-hate-speech-in-its-ads.html [https://perma.cc/723Y-VAMK].

20. *See* Rodriguez, *supra* note 19.

21. *See, e.g.*, Newton, *supra* note 18 (stating YouTube promised to "reconsider all of its harassment policies" in response to public outcry for not removing homophobic content); Rodriguez, *supra* note 19 (stating Facebook announced that it would ban

In fact, the policy decisions that social platforms make are vague, opaque, and often appear to do little more than pay lip service to their ideals of corporate social responsibility. Simply put: they just do not do enough. For example, although Facebook has repeatedly assured its users that the platform does not tolerate online harassment, [22] an independent audit recently found that Facebook has allowed hate speech and disinformation to thrive. [23] Although Twitter emphasizes what progress it has made in addressing these issues, it acknowledges that online abuse remains a problem on its platform.[24] Likewise, YouTube highlights its willingness to make policy adjustments as needed to combat abuse, yet acknowledges that it has not done enough.[25] Even the recent announcements about efforts to stop the spread of false information concerning the coronavirus pandemic may turn out to be little more than window dressing, given that a recent report found "some of the most dangerous falsehoods had received hundreds of thousands of views." [26]

One reason social platforms struggle is because content moderation is difficult. Given the sheer number of new content posted in one minute on any given platform, at least part of the content regulation must be automated. Yet language is incredibly complicated, personal, and context dependent, which limits the algorithms' abilities to differentiate between permissible and problematic posts.[27] Another reason platforms struggle is because they have been unwilling to draw clear lines regarding what content violates their policies and consistently and transparently enforcing them. [28]

---

ads containing hate speech only after almost 100 advertisers began boycotting Facebook for its failure to police hate speech).

22. Elizabeth Cohen, *She Was Called the N-Word and Given Instructions to Slit Her Wrists. What Did Facebook Do?*, CNN HEALTH, https://www.cnn.com/2019/11/01/health/facebook-harassment-eprise/index.html (Nov. 1, 2019, 2:06 PM), [https://perma.cc/W8CD-9B8U].

23. Mike Isaac, *Facebook's Decisions Were 'Setbacks for Civil Rights,' Audit Finds*, N.Y. TIMES (Aug. 3, 2020), https://www.nytimes.com/2020/07/08/technology/facebook-civil-rights-audit.html [https://perma.cc/8X46-WRRY].

24. Katy Minshall, #*SaferInternetDay 2020: Creating a Better Internet for All*, TWITTER BLOG (Feb. 11, 2020), https://blog.twitter.com/en_us/topics/events/2020/safer-internet-day-2020-creating-a-better-internet-for-all.html [https://perma.cc/5PK5-3H44].

25. Julia Alexander, *YouTube's 'Creator-on-Creator' Harassment Policy Could Affect Commentary Videos*, THE VERGE (Aug. 5, 2019, 12:02 PM), https://www.theverge.com/2019/8/5/20754620/youtube-commentary-harassment-creators-bullying-susan-wojcicki [https://perma.cc/P9QW-X6Y7].

26. Joe Tidy, *Coronavirus: Facebook Alters Virus Action After Damning Misinformation Report*, BBC (Apr. 16, 2020), https://www.bbc.com/news/technology-52309094 [https://perma.cc/4DMZ-KAXL].

27. Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y 1, 5 (Feb. 28, 2020), https://doi.org/10.1177/2053951719897945 [https://perma.cc/8M6L-HLPA].

28. Sarah Jeong, *Turns Out Facebook Moderation Sucks Because Its Guidelines Suck*, THE VERGE (Apr. 24, 2018, 4:51 PM), https://www.theverge.com/2018/4/24/

Ultimately, social platforms engage in self-regulation privately and selectively, and speech harms continue to thrive online. In response, citizens and lawmakers have consistently called for the federal government to regulate social media content on some level. The suggestion has gained momentum, and while most social platforms have balked at the idea,[29] some have specifically sought it out.[30]

The thorn in any effort to regulate content on social platforms is, of course, the First Amendment. With this in mind, a variety of different options have been suggested to ameliorate harmful content without running afoul of the Constitution. Several legislators and commentators have suggested significantly amending or altogether eliminating the protections of section 230 of the CDA.[31] Others have explored the option of wholesale government regulation of social platforms as public utilities.[32]

Regardless of the form content moderation takes, the growing cry of citizens, lawmakers, and even some social platforms for federal government oversight creates a likelihood of government intervention within the next few years.[33] The risk is that this momentum for change will result in a rush to regulate without proper consideration of the true costs and benefits.[34] The enthusiasm across party lines for re-

17276794/facebook-moderation-guidelines-community-standards-nudity-hate-speech [https://perma.cc/435F-7RMW]; Ahmad Sultan, *We Need Real Transparency About Hate on Social Media*, ANTI-DEFAMATION LEAGUE (Jan. 2, 2019), https://www.adl.org/blog/we-need-real-transparency-about-hate-on-social-media [https://perma.cc/BP8H-69BX].

29. Natasha Tusikov & Blayne Haggart, *It's Time for a New Way to Regulate Social Media Platforms*, THE CONVERSATION (Jan. 16, 2019, 6:48 PM), https://theconversation.com/its-time-for-a-new-way-to-regulate-social-media-platforms-109413 [https://perma.cc/L66G-NWDC].

30. *See* Andrew Hutchinson, *Facebook Publishes New Whitepaper on Standardized Online Content Regulation*, SOC. MEDIA TODAY (Feb. 18, 2020), https://www.socialmediatoday.com/news/facebook-publishes-new-whitepaper-on-standardized-online-content-regulation/572416/ [https://perma.cc/23RE-5FJ4]; Mark Zuckerberg, Opinion, *The Internet Needs New Rules. Let's Start in These Four Areas.*, WASH. POST (Mar. 30, 2019, 2:00 PM), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html [https://perma.cc/ZU98-WCB2].

31. *See, e.g.*, Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017); Anthony Ciolli, *Chilling Effects: The Communications Decency Act and the Online Marketplace of Ideas*, 63 U. MIA. L. REV. 137 (2008); David Lukmire, *Can the Courts Tame the Communications Decency Act?: The Reverberations of* Zeran v. America Online, 66 N.Y.U. ANN. SURV. AM. L. 371 (2010); Benjamin Volpe, *From Innovation to Abuse: Does the Internet Still Need Section 230 Immunity?*, 68 CATH. U. L. REV. 597 (2019).

32. *See, e.g.*, K. Sabeel Rahman, *Regulating Informational Infrastructure: Internet Platforms as the New Public Utilities*, 2 GEO. L. TECH. REV. 234 (2018).

33. Ryan Tracy, *Social Media's Liability Shield Is Under Assault*, WALL ST. J. (Nov. 26, 2020, 10:00 AM), https://www.wsj.com/articles/social-medias-liability-shield-is-under-assault-11606402800 [https://perma.cc/T2Z6-WESS].

34. *See* Steven Greenhut, *The Bipartisan Push to Gut Section 230 Will Suppress Online Speech*, REASON.COM (Dec. 18, 2020, 8:00 AM), https://reason.com/2020/12/18/

duced section 230 protections is particularly worrisome because the issues of content regulation and section 230 reform are distinct policy issues that all too often become ensnared in debate.

This Article acknowledges the need for some level of content regulation for social platforms and explores three possible avenues for its execution: private governance via platform self-regulation (the current system), government regulation, and industry self-regulation.[35] As this Article will demonstrate, an industry-wide governance model is the optimal solution to reduce harmful speech on social media without hindering the free exchange of ideas.

Part II outlines the current structure for permissible regulation with a particular focus on section 230 of the CDA, since efforts to regulate content online tend to center around this law. Part III addresses the landscape of government regulation and endeavors to untangle the conversation about section 230 from the debate about content regulation. Part IV outlines systems of self-regulation, both at the platform and industry levels. Part V concludes with a recommendation for a system that addresses concerns of citizens and lawmakers who want to reduce harms on social media while also balancing the interests of social platforms and robust speech rights.

II. The Lay of the Land: The Current Structure of Content Regulation on Social Platforms

Strong First Amendment protections apply across all media, including, of course, the Internet.[36] In order to comply with the First Amendment, wholesale regulation of mass communication has traditionally depended on a "medium-specific" approach by the courts.[37] The broadcast media, for example, have customarily been subject to the most government regulation and oversight.[38] The Court has upheld these regulations based on the invasive nature of the broadcast

the-bipartisan-push-to-gut-section-230-will-suppress-online-communication/ [https://perma.cc/AY84-HF68].

35. This Article is necessarily limited to examining methods of regulation based on the substance of communications and does not address a separate area under discussion for more regulation: social platforms and data privacy.

36. Ashley Fuchs, *Proceed with Caution: Why Curtailing Section 230 Immunity Is Not the Solution to Social Media Regulation*, Univ. Penn.: Ctr. Ethics & Rule L. (Sept. 28, 2020), https://www.law.upenn.edu/live/news/10511-proceed-with-caution-why-curtailing-section-230 [https://perma.cc/869D-CCEW].

37. Reno v. ACLU, 521 U.S. 844, 863 (1997) (quoting ACLU v. Reno, 929 F. Supp. 824, 873 (E.D. Pa. 1996)).

38. Consider FCC regulations of indecent speech on the broadcast networks, 18 U.S.C. § 1464, or even the now-defunct Fairness Doctrine. *See, e.g.*, Red Lion Broad. Co. v. FCC, 395 U.S. 367, 370–71, 375 (1969) (upholding constitutionality of the fairness doctrine); Maria Fontenot & Michael T. Martínez, *FCC's Indecency Regulation: A Comparative Analysis of Broadcast and Online Media*, 26 UCLA Ent. L. Rev. 59, 72 (2019) (discussing differences between Internet and broadcast regulations); Khaldoun Shobaki, *Speech Restraints for Converged Media*, 52 UCLA L. Rev. 333, 343 (2004) (discussing differences in regulation of media forms).

airwaves, their status as scarce expressive commodities, and the public's entitlement to receive suitable access to ideas.[39] At the other end of the spectrum, the print media is neither scarce nor invasive, and as such, have a long history of robust protection under the First Amendment.[40] Thus, courts have consistently bristled at congressional attempts to regulate print media.[41]

Like print media, the Internet is neither invasive nor scarce.[42] When it emerged as a new medium of mass communication, the Supreme Court opted to treat it in the same vein as print media—with strong First Amendment protection.[43] The upshot is that online speech is largely free from regulation based on the substance of the message being communicated, regardless of the online platform or the content creator's identity.[44]

The broad protection for online speech means that few federal or state guidelines exist that direct social platforms to police or remove speech based on its content, leaving platforms in control over the vast majority of content that resides on their networks.[45] These platforms have benefited enormously from this scheme and likely would not exist without it.[46] Their business models rely on the ability of millions of users to create and upload content without direction or oversight, and advertisers who count on constant user engagement with that content.[47] A regulatory framework permitting the government to prohibit certain messaging would create a technological barrier to optimizing this business model: platforms would need to have the means to filter, separate, and potentially block content based upon the message communicated.[48] As will be discussed further in Part IV, even the most sophisticated algorithms struggle with content decisions, making a

39. *Red Lion Broad. Co.*, 395 U.S. at 387–90; FCC v. Pacifica Found., 438 U.S. 726, 731 n.2 (1978); Turner Broad. Sys., Inc. v. FCC (*Turner II*), 520 U.S. 180, 185 (1997); Turner Broad. Sys., Inc. v. FCC (*Turner I*), 512 U.S. 622, 637–39 (1994).

40. *See* Robert W. McChesney, *Freedom of the Press for Whom? The Question To Be Answered in Our Critical Juncture*, 35 HOFSTRA L. REV. 1433, 1444, 1446 (2007).

41. Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 256 (1974).

42. *Reno*, 521 U.S. at 869–70.

43. *Id.* at 863 (quoting ACLU v. Reno, 929 F. Supp. 824, 883 (E.D. Pa. 1996)).

44. Acknowledgement of broad First Amendment protection for Internet speech has frustrated several attempts at government regulation of content, such as those aimed at restricting sexually explicit material that could be harmful to minors. *See, e.g.*, *id.* at 844 (striking down anti-indecency provisions of the CDA); Ashcroft v. Free Speech Coal., 535 U.S. 234 (2002) (upholding the injunction on enforcement of the Child Online Protection Act ("COPA")).

45. Kate Conger, *Facebook, Google and Twitter C.E.O.s Return to Washington to Defend Their Content Moderation*, N.Y. TIMES (Oct. 28, 2020), https://www.nytimes.com/2020/10/28/technology/facebook-google-and-twitter-ceos-return-to-washington-to-defend-their-content-moderation.html [https://perma.cc/X7F4-VD7N].

46. *See id.*

47. Tracy, *supra* note 33.

48. *See* Daisuke Wakabayashi, *Legal Shield for Social Media Is Targeted by Lawmakers*, N.Y. TIMES (Dec. 15, 2020), https://www.nytimes.com/2020/05/28/business/section-230-internet-speech.html [https://perma.cc/HR2U-S8NP].

complete technological solution improbable. Given the sheer amount of content posted, relying on human moderators to filter user posts and fill this gap would create another massive burden.[49]

It is not the case, however, that there is *no* speech regulation online or on social platforms. While most attempts at regulatory action have not been successful,[50] there exist discrete areas where regulation is permissible. One example is a federal law that requires Internet sites to remove images of child pornography.[51] Another is the Digital Millennium Copyright Act ("DMCA"), passed in 1998 to limit the liability of internet service providers ("ISP") for copyright infringement by their users.[52] The DMCA has a notice-and-takedown provision that, to merit safe harbor from copyright infringement charges, requires ISPs to remove infringing material once they are on notice of such material's existence.[53]

Importantly, platforms also benefit from a federal law that immunizes them from liability for the content their users post.[54] Arguably the most important law for online speech, section 230 protects ISPs from liability for content posted on their sites by third parties.[55] This law enables social platforms to function by allowing their users to freely create and upload content with little risk of liability for the publisher of that content—the platform. It is also an increasingly controversial law.

A. *The Benefit of Section 230*

The reason section 230 offers such robust protection from liability for social platforms is because in crafting the law, Congress made the policy decision to treat online publishers differently than print publishers.[56] Essentially, Congress took the Court's decision to treat the internet like print media one step further and protected internet publishers more than their print counterparts.[57] The law was written

49. *Id.*

50. *See* Nicholas P. Dickerson, Comment, *What Makes the Internet So Special? And Why, Where, How, and by Whom Should Its Content Be Regulated?*, 46 HOUS. L. REV. 61, 78 (2009) (citing Ashcroft v. ACLU, 542 U.S. 673 (2004)). Attempts to restrict sexually explicit material that could be harmful to minors have been overwhelmingly unsuccessful. *See, e.g.*, *Ashcroft*, 542 U.S. at 673; Martha McCarthy, *The Continuing Saga of Internet Censorship: The Child Online Protection Act*, 2005 BYU EDUC. & L.J. 83, 89 (2005).

51. 18 U.S.C. §§ 2251–2252A.

52. 17 U.S.C. § 512(g)(1)–(4).

53. Amanda Reid, *Considering Fair Use: DMCA's Take Down & Repeat Infringers Policies*, 24 COMMC'N L. & POL'Y 101, 102 (2019).

54. 47 U.S.C. § 230 (c)(1), (f)(2). As Prof. Jeff Kossef has written, section 230 is best understood as *The Twenty-Six Words That Created the Internet*. JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

55. 47 U.S.C. § 230(c).

56. *See* Batzel v. Smith, 333 F.3d 1018, 1026–28 (9th Cir. 2003) (citing Blumenthal v. Drudge, 992 F. Supp. 44, 49 (D.D.C. 1998)).

57. *Id.*

before the emergence of social platforms and applies to immunize any "provider or user of an interactive computer service" when that "provider or user" republishes content created by another user.[58] The law defines the term *user* broadly, and applies it "simply to anyone using an interactive computer service."[59] The protection, too, is broad. Section 230 shields providers from liability for their decisions to moderate content, or to transmit content as is, without moderation.[60]

Thus, a range of interactive computer service providers—particularly online services that rely on publishing third-party content, such as social platforms—benefit from section 230.[61] Section 230 has immunized Facebook, Google (YouTube), Yahoo!, and others from liability stemming from third-party content, even when the platform knew about, tried to block, removed, or policed the content.[62] Courts have applied this immunity widely, encompassing claims for defamation, negligence, intentional infliction of emotional distress, privacy, terrorism support, and more.[63]

In passing section 230, Congress elected to protect ISPs from intermediary liability for the activity of their users in order to promote the uninhibited flow of ideas throughout the internet without government interference. However, this decision has also enabled harmful speech to thrive online.[64] As explained in *Zeran v. America Online, Inc.*, the seminal case addressing the reach of section 230:

> Congress recognized the threat that tort-based lawsuits pose to freedom of speech in the new and burgeoning Internet medium. The imposition of tort liability on service providers for the communications of others represented, for Congress, simply another form

58. 47 U.S.C. § 230(c).

59. Barrett v. Rosenthal, 146 P.3d 510, 515 (Cal. 2006).

60. 47 U.S.C. § 230(c).

61. *Id.*

62. *See generally* Fields v. Twitter, Inc., 217 F. Supp. 3d 1116, 1118 (N.D. Cal. 2016), *aff'd*, 881 F.3d 739 (9th Cir. 2018); Klayman v. Zuckerberg, 753 F.3d 1354, 1355 (D.C. Cir. 2014); Goddard v. Google, Inc., 640 F. Supp. 2d 1193, 1195–96 (N.D. Cal. 2009); Gentry v. eBay, Inc., 121 Cal. Rptr. 2d 703, 706 (Cal. Ct. App. 2002); Barnes v. Yahoo!, Inc. (*Barnes II*), 570 F.3d 1096, 1102–03, 1105–06 (9th Cir. 2009); Carafano v. Metrosplash.com, Inc., 339 F.3d 1119, 1119 (9th Cir. 2003); Dart v. Craigslist, Inc., 665 F. Supp. 2d 961, 961 (N.D. Ill. 2009).

63. *See, e.g.*, Batzel v. Smith, 333 F.3d 1018, 1026–27 (9th Cir. 2003) (involving a claim for defamation); Ben Ezra, Weinstein, & Co. v. Am. Online, Inc., 206 F.3d 980, 983–84 (10th Cir. 2000) (involving a claim for defamation and negligence); Zeran v. Am. Online, Inc., 129 F.3d 327, 330, 332 (4th Cir. 1997) (involving negligence claims); Beyond Sys., Inc. v. Keynetics, Inc., 422 F. Supp. 2d 523, 525, 536 (D. Md. 2006) (involving a claim under Maryland Commercial Electronic Mail Act); Doe v. Bates, No. 5:05-CV-91-DF-CMC, 2006 WL 3813758, at *2–3, *5–6 (E.D. Tex. Dec. 27, 2006) (involving claims of negligence, negligence per se, intentional infliction of emotional distress, invasion of privacy, civil conspiracy, and distribution of child pornography); Barnes v. Yahoo!, Inc., No. Civ. 05–296–AA, 2005 WL 3005602, at *1 (D. Or. Nov. 8, 2005) (involving a negligence claim resulting in personal injury).

64. *Zeran*, 129 F.3d at 330–31. Congress's decision in this light gives rise to a valid argument that the Internet is the least-regulated form of media, even behind print.

> of intrusive government regulation of speech. Section 230 was enacted, in part, to maintain the robust nature of Internet communication and, accordingly, to keep government interference in the medium to a minimum. In specific statutory findings, Congress recognized the Internet and interactive computer services as offering "a forum for a true diversity of political discourse, unique opportunities for cultural development, and myriad avenues for intellectual activity." [47 U.S.C.] § 230(a)(3). It also found that the Internet and interactive computer services "have flourished, to the benefit of all Americans, *with a minimum of government regulation*." *Id.* § 230(a)(4). Congress further stated that it is "the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, *unfettered by Federal or State regulation*." *Id.* § 230(b)(2).[65]

The goal of Congress's decision in enacting section 230 was that Internet companies would be encouraged to develop platforms that relied almost entirely on user-generated content without fear of liability for the content users posted. Without section 230, "the [potential] liability that would arise from allowing users to freely exchange information with one another, at this [large] scale, would have been astronomical" and could very well have prevented investors from supporting social platforms.[66]

Considering the sheer volume of material that is posted to many websites on any given day, policing all of the content would prove impossible. As of October 2020, Facebook reported 2.74 billion active users who log in to their accounts at least once a month.[67] Each day, users create a cumulative total of 4.3 billion posts and upload 8 billion hours of video content.[68] On Twitter, roughly 6,000 new tweets emerge each second, totaling 350,000 tweets per minute, 500 million per day, and 200 billion per year.[69] On YouTube, 500 hours of content

65. *Id.*

66. David Post, Opinion, *A Bit of Internet History, or How Two Members of Congress Helped Create a Trillion or So Dollars of Value*, WASH. POST (Aug. 27, 2015, 12:05 PM), https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/08/27/a-bit-of-internet-history-or-how-two-members-of-congress-helped-create-a-trillion-or-so-dollars-of-value [https://perma.cc/RX28-6WJL].

67. Press Release, Facebook, Facebook Reps. Third Quarter 2020 Results (Oct. 29, 2020) (available at https://s21.q4cdn.com/399680738/files/doc_news/Facebook-Reports-Third-Quarter-2020-Results-2020.pdf [https://perma.cc/3PFZ-FRHD]); *see also* Press Release, Facebook, Facebook Reps. First Quarter 2020 Results (Apr. 29, 2020) (available at https://s21.q4cdn.com/399680738/files/doc_news/Facebook-Reports-First-Quarter-2020-Results-2020.pdf [https://perma.cc/6N7D-KXKT]).

68. *Facebook Statistics and Facts*, MARKET.US, https://market.us/statistics/social-media/facebook/ [https://perma.cc/6E5K-YJ6W].

69. *Twitter Usage Statistics*, INTERNET LIVE STAT., https://www.internetlivestats.com/twitter-statistics/ [https://perma.cc/77DD-T66V].

are uploaded every minute.[70] Given this volume, it is easy to understand the value of section 230 for these platforms: intermediary liability would cripple their business models.

But the benefits of section 230 extend beyond the liability shield for large social platforms. As an initial matter, it applies to all *users* of interactive services, which necessarily includes individuals,[71] corporations,[72] non-profit organizations,[73] and more. It is easy to assume that the greatest beneficiaries of the law are large social platforms—and there may be some truth to that—but start-ups and small companies also gain an advantage from this framework. Section 230 "deters frivolous and costly lawsuits, and it speeds up resolution when such lawsuits are brought,"[74] which enables small start-up companies to find a foothold in the online space and "encourage[s] the next generation of start-up businesses aspiring to disrupt the current Internet incumbents."[75]

Despite the businesses—large and small—that benefit from section 230, the ultimate beneficiaries are the *users* of interactive computer services. Without section 230's protections, users would not find an online space to quickly create and share thoughts, photos, and videos, and view those posted by others.[76] The ability to freely comment on posts created by others would be stifled, as would the ability to write—or read—product reviews.[77] Further, these activities would still be possible with significant content moderation that would lead to a time lag between initial posting and ultimate online experience. This lag would fundamentally change the way users utilize and rely on online space.

Of course, immunity under section 230 is not unlimited. Users who create harmful content expose themselves to liability for that content.

---

70. Mansoor Iqbal, *YouTube Revenue and Usage Statistics (2020)*, BUS. APPS (Nov. 17, 2020), https://www.businessofapps.com/data/youtube-statistics/ [https://perma.cc/6S92-4SGL].

71. *See, e.g.*, Mitan v. A. Neumann & Assocs., LLC, Civ. No. 08-6154, 2010 WL 4782771, at *17–18 (D.N.J. Nov. 17, 2010) (dismissing claim based on defamatory contents in a forwarded email); Barrett v. Rosenthal, 146 P.3d 510, 528–29 (Cal. 2006) (holding that CDA immunity extended to individuals who republished defamatory statements via the internet originally made by others in email and internet postings).

72. *See, e.g.*, Force v. Facebook, Inc., 934 F.3d 53, 57 (2d Cir. 2019), *cert. denied*, 140 S. Ct. 2761 (2020).

73. *See, e.g.*, Nat'l Ass'n of the Deaf v. Harvard Univ., 377 F. Supp. 3d 49, 69 (D. Mass. 2019).

74. Letter from forty-six Acads. to Members of Cong. (Mar. 9, 2020) (on file at https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=3164&context=historical [https://perma.cc/2PN4-VPWS]).

75. *Id.*

76. Alina Selyukh, *Section 230: A Key Legal Shield for Facebook, Google is About to Change*, NPR (Mar. 21, 2018, 5:11 AM), https://www.npr.org/sections/alltechconsidered/2018/03/21/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change [https://perma.cc/P8BK-KVLT].

77. *Id.*

This applies to social platforms, too—when any interactive computer service develops content, or "contributes materially to the alleged illegality of the conduct," it is legally responsible for that content.[78] In other words, platforms can be held liable for any content they *create* on their own or *cause* to exist.

There are five specific statutory exceptions to immunity, including those sounding in (1) federal criminal law, (2) intellectual property law, (3) state law, (4) communications privacy law, and (5) sex trafficking law.[79] The fifth exception is a recent addition.[80] In 2018, Congress enacted the Fight Online Sex Trafficking Act ("FOSTA"), clarifying that section 230 immunity will not offer protection for several sex trafficking offenses, regardless of whether the service provider materially contributed to the unlawful conduct.[81]

B. *Efforts to Remove the Shield of Section 230*

Section 230's breadth has made it an easy target for those hungry for changes in the way social platforms curate—or do not curate—their platforms. Commentators and legislators from across the political spectrum have called for amending or otherwise reducing the protections afforded by section 230.[82] These challenges can be categorized into two basic groups: (1) those based on ideals of viewpoint neutrality for platforms and (2) concerns that platforms do too little to remove harmful content, such as cyber bullying,[83] sexual harassment,[84] cyberstalking,[85] nonconsensual pornography,[86] and defamation.[87]

78. Fair Hous. Council v. Roommates.com, LLC, 521 F.3d 1157, 1168 (9th Cir. 2008) (en banc) (holding that Roommates.com was immune for claims arising from the content that users provided but not from those arising from required questions it asked users); FTC v. Accusearch Inc., 570 F.3d 1187, 1199 (10th Cir. 2009) (denying section 230 immunity where website "was responsible for the development of that content—for the conversion of the legally protected records from confidential material to publicly exposed information").

79. 47 U.S.C. § 230(e).

80. *Id.*

81. Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164, 132 Stat. 1253 (2018).

82. Nandita Bose & Raphael Satter, *Should Facebook, Google be Liable for User Posts? Asks U.S. Attorney General Barr*, REUTERS (Feb. 19, 2020, 9:51 AM), https://www.reuters.com/article/us-internet-regulation-justice/should-facebook-google-be-liable-for-user-posts-asks-u-s-attorney-general-barr-idUSKBN20D26S [https://perma.cc/UAG3-JLMM].

83. EMILY BAZELTON, STICKS AND STONES: DEFEATING THE CULTURE OF BULLYING AND REDISCOVERING THE POWER OF CHARACTER AND EMPATHY (2013).

84. Mary Anne Franks, *Sexual Harassment 2.0*, 71 MD. L. REV. 655, 657 (2012).

85. *See* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 171–72 (Harvard Univ. Press 2014).

86. *Id.*

87. Lee K. Royster, *Fake News: Potential Solutions to the Online Epidemic*, 96 N.C. L. REV. 270, 275 (2017); Neville L. Johnson, Douglas L. Johnson, Paul Tweed &

1. Calls to Leverage Immunity in Exchange for Viewpoint Neutrality

Lawmakers on both sides of the aisle have criticized section 230 and have pushed for significant changes to it.[88] However, conservatives particularly favor calling for changes based on the ideals of viewpoint neutrality for platforms.[89] The essence of the argument is that social platforms "censor opinions with which they disagree," which is particularly problematic given that "these platforms function in many ways as a 21st century equivalent of the public square."[90]

Additionally, as private actors, the First Amendment does not constrain social platforms as it would against government actors.[91] Courts have held that private online service providers are *not* state actors for First Amendment purposes,[92] which means the First Amendment has no direct role to play in regulating the content policies and practices of social media companies.[93] In spite of this, users have still sued social platforms on the basis of alleged conservative discrimination,[94] and conservative lawmakers continue to argue that section 230 protections ought to be removed for platforms that demonstrate such an alleged bias.[95]

Rodney A. Smolla, *Defamation and Invasion of Privacy in the Internet Age*, 25 Sw. J. Int'l 9, 39 (2019).

88. Marguerite Reardon, *What's Section 230? The Social Media Law That's Clogging up the Stimulus Talks*, CNET (Dec. 30, 2020, 8:16 AM), https://www.cnet.com/news/whats-section-230-the-social-media-law-thats-clogging-up-the-stimulus-talks/ [https://perma.cc/EJ3N-DE5Q].

89. *Id.*

90. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020).

91. *See generally* Brown & Peters, *supra* note 4, at 540.

92. *See, e.g.*, Prager Univ. v. Google LLC, 951 F.3d 991, 999 (9th Cir. 2020) (holding that the state action doctrine precluded constitutional scrutiny of private entity's content moderation); Wilson v. Twitter, No. 3:20-cv-00054, 2020 WL 4353686, at *5 (S.D.W.V. May 1, 2020) ("That private social media companies now host platforms which imitate the functions of public forums—in many respects more effectively than the traditional public forums of government-owned sidewalks, streets, and public parks—does not mean that the entities are state-actors for the purposes of the First Amendment."); Tulsi Now, Inc. v. Google, LLC, No. 2:19-cv-06444-SVW-RAO, 2020 WL 4353686, at *2–3 (C.D. Cal. Mar. 3, 2020); Fed. Agency of News LLC v. Facebook, Inc., 432 F. Supp. 3d 1107, 1126–27 (N.D. Cal. 2020).

93. Brown & Peters, *supra* note 4, at 540.

94. *See* Freedom Watch, Inc. v. Google, Inc., 368 F. Supp. 3d 30, 41 (D.D.C. 2019) ("[T]he Amended Complaint focuses on the Platforms' alleged suppression of conservative political content. It details, for instance, the seemingly disparate treatment of conservative news publishers on Facebook and of conservative commentators on Twitter. But while selective censorship of the kind alleged by the Plaintiffs may be antithetical to the American tradition of freedom of speech, *it is not actionable under the First Amendment unless perpetrated by a state actor.*" (emphasis added) (citation omitted)).

95. *See, e.g.*, Cale Guthrie Weissman, *Ted Cruz Made It Clear He Supports Repealing Tech Platforms' Safe Harbor*, Fast Co. (Oct. 17, 2018), https://www.fastcompany.com/90252598/ted-cruz-made-it-clear-he-supports-repealing-tech-platforms-safe-harbor [https://perma.cc/B3SY-LBAD].

One recent attempt to translate this belief into action occurred in May 2020, when President Trump issued an executive order purporting to limit section 230's protections.[96] The move was easily interpreted as retaliatory: it came days after Twitter decided to add a fact-check label to two of the President's arguably false tweets about mail-in voting. The executive order tasked the Federal Communications Commission ("FCC") with redefining when section 230 applies and the Federal Trade Commission ("FTC") with ensuring that social platforms adhere to their own terms and conditions.[97] Under the order, the Department of Justice ("DOJ") would also review a list of platforms and determine whether they impose "viewpoint-based speech restrictions" and are therefore "problematic vehicles for government speech."[98] If so, the DOJ would restrict or curtail government ad-spending on those platforms.[99]

The FCC and the DOJ have already taken action. In June 2020, the DOJ submitted its recommendations for amending section 230 pursuant to Trump's executive order. The proposal identified four areas that were ripe for reform, including amending section 230 to incentivize platforms to reduce illegal content on their sites, clarifying the federal government's ability to enforce claims on behalf of citizens, increasing competition among the social platforms, and increasing disclosure and transparency in content moderation processes.[100] In October 2020, FCC Chairman Ajit Pai announced that the agency would "move forward with a rulemaking to clarify [section 230's] meaning."[101]

However, the executive order is constitutionally problematic because it is an effort to both force and limit the speech of social platforms. The First Amendment protects both of these activities, as platforms have the freedom to determine what content appears on their sites or refrain from speaking at all. Thus, requiring a social platform to host content that it would otherwise prohibit or limit is a clear violation of its First Amendment rights.[102] The executive order suffers from other legal deficiencies, most notably that it is not legally sufficient to amend an existing statute, particularly one with a twenty-five-year history of judicial interpretations inconsistent with the order's

96. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020).
97. *Id.*
98. *Id.*
99. *Id.*
100. *Section 230 – Nurturing Innovation or Fostering Unaccountability?*, U.S. DEP'T JUST. 1, 7 (June 2020), https://www.justice.gov/file/1286331/download [https://perma.cc/UWT8-68VP].
101. Statement of Chairman Pai on Section 230, FCC (October 2020), https://docs.fcc.gov/public/attachments/DOC-367567A1.pdf [https://perma.cc/6N9D-T4RR].
102. Brown & Peters, *supra* note 4.

terms.[103] But the fact that the order has little support in the law was probably beside the point, because the ultimate goal was instead to intimidate those that might limit what Mr. Trump says, even when it is blatantly false or harmful. It was to send a message to social platforms: any effort to limit or frame the president's speech, even when false and potentially damaging, will be met with aggressive legal action.

Other lawmakers have also tried to limit section 230's sweeping protections for social platforms on the basis that platforms make overt efforts to censor conservative speech.[104] Senator Ted Cruz has been a vocal opponent of section 230, arguing that platforms should be content-neutral, and has asked the administration to modify proposed trade deals to remove language that offers immunity from liability similar to section 230.[105] During congressional hearings in August 2020, Representatives Jim Sensenbrenner and Jim Steube accused Facebook of filtering out conservative speech on its platform.[106]

Senator Josh Hawley has introduced at least three bills since 2019 to amend section 230 to remove immunity unless tech companies prove their algorithms and content-removal practices are politically neutral.[107] One of Senator Hawley's proposals, the Ending Support for Internet Censorship Act, is part of a class of proposed legislation that aims to tackle perceived censorship of conservative speech online.[108]

---

103. *See* Youngstown Sheet & Tube Co. v. Sawyer, 343 U.S. 579, 587 (1952) ("[T]he President's power to see that the laws are faithfully executed refutes the idea that he is to be a lawmaker.").

104. There is partisan support for this belief. *How Can Social Media Firms Tackle Hate Speech?*, KNOWLEDGE @ WHARTON (Sep. 22, 2018), https://knowledge.wharton.upenn.edu/article/can-social-media-firms-tackle-hate-speech/ [https://perma.cc/5JGT-KDES] (noting that a Pew Research poll from June 2018 found that "many Americans perceive social media as playing an active role in censorship. When asked whether they think it likely that social platforms actively censor political views that those companies find objectionable, 72% of respondents . . . said yes. Republicans were especially inclined to think so: 85% of Republicans and Republican-leaning independents said it was likely that social media sites intentionally censor political viewpoints, with 54% saying it was very likely, found the Pew survey of 4,594 U.S. adults.").

105. Letter from Ted Cruz, Texas Senator, to Robert Lighthizer, United States Trade Representative (Nov. 1, 2019), https://www.cruz.senate.gov/files/documents/2019.11.01_USTR%20Sec%20230%20LTR.pdf [https://perma.cc/8V8Q-NFPR]; Elliot Harmon, *No, Section 230 Does Not Require Platforms to Be "Neutral"*, EFF (April 12, 2018), https://www.eff.org/deeplinks/2018/04/no-section-230-does-not-require-platforms-be-neutral [https://perma.cc/E344-JSU5]; Weissman, *supra* note 95.

106. *Online Platforms and Market Power: Hearing Before the Subcomm. on the Judiciary*, 116th Cong. (2020).

107. *See, e.g.*, *Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies*, JOSH HAWLEY U.S. SENATOR FOR MO. (June 19, 2019), https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies [https://perma.cc/U9UM-NBR8].

108. Zoe Bedell & John Major, *What's Next for Section 230? A Roundup of Proposals*, LAWFARE: SOC. MEDIA (July 29, 2020, 9:01 AM), https://www.lawfareblog.com/whats-next-section-230-roundup-proposals [https://perma.cc/BZ8Y-6G75].

Hawley's proposal focuses on revoking immunity for social platforms that are unable to demonstrate their content moderation practices are politically neutral.[109] If passed, the Ending Support for Internet Censorship Act would force tech giants to apply for immunity through the FTC every two years.[110] An applicant would have to prove that it employs politically neutral content moderation practices by clear and convincing evidence, and the FTC would then take a vote to grant immunity if approved by a supermajority.[111]

Senators Hawley and Marco Rubio have also proposed the Limiting Section 230 Immunity to Good Samaritans Act.[112] This proposal would make a social network's immunity under section 230 contingent upon the network's contractual commitment to using "good faith" practices when making content moderation decisions.[113] This good faith requirement would bar tech companies from enforcing their terms unevenly based on perceived political bias.[114]

Finally, Senator Hawley also introduced the BAD ADS Act, which, unlike his other proposals, does not attempt to impose political neutrality on tech companies and their content moderation processes.[115] Instead, this proposal aims to prevent social networks from using behavioral advertising practices on their platforms.[116] To achieve this objective, the proposal revokes section 230 immunity for thirty days whenever a social network uses behavioral advertising practices on its site.[117] Such practices consist of collecting user data to create advertising profiles based on a user's personal demographics and online activity then using that data to generate user-specific advertisements.[118]

---

109. *Id.*

110. Mary Catherine Wellons, *GOP Senator Introduces a Bill That Would Blow Up Business Models for Facebook, YouTube and Other Tech Giants*, CNBC: TECH (Jun. 19, 2019, 8:27 AM), https://www.cnbc.com/2019/06/18/sen-hawley-bill-would-revoke-cda-section-230-for-large-tech-companies.html [https://perma.cc/L78Y-LUCT].

111. *Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies*, JOSHUA HAWLEY U.S. SENATOR FOR MO. (Jun. 19, 2019), https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies [https://perma.cc/8P8R-2P8Y].

112. Lauren Feiner, *GOP Sen. Hawley Unveils His Latest Attack on Tech's Liability Shield in New Bill*, CNBC: TECH (Jun. 17, 2020, 10:26 AM), https://www.cnbc.com/2020/06/17/gop-sen-hawley-unveils-latest-attack-on-techs-liability-shield.html [HTTPS://PERMA.CC/5FYB-WBGX].

113. *Id.*

114. Press Release, Marco Rubio, U.S. Sen. for Fla., Rubio, Hawley Announce Bill Empowering Americans to Hold Big Tech Companies Accountable for Acting in Bad Faith (June 17, 2020), https://www.rubio.senate.gov/public/index.cfm/press-releases?ContentRecord_id=47276D77-62D6-4E04-9FA2-1CD761179B90#:~:text=background%3A%20The%20Limiting%20Section%20230,a%20duty%20of%20good%20faith [https://perma.cc/V8UW-YFJ9].

115. *See* Bedell & Major, *supra* note 108.

116. *Id.*

117. *Id.*

118. *Id.*

The Online Freedom and Viewpoint Diversity Act ("OFVDA") is conservative lawmakers' most recent attempt to condition section 230 immunity upon politically neutral content moderation practices.[119] In order to accomplish this, the OFVDA would remove the catchall language in section 230(c)(2)(A) that currently affords social platforms civil immunity when they remove user content that is "otherwise objectionable."[120] In addition, the OFVDA would amend section 230(c)(2)(A) to specifically grant immunity when platforms censor content that is "unlawful," "promotes terrorism," or "promotes self-harm."[121]

In practice, these modifications would force platforms to surrender their independent discretion in exchange for immunity. Thus, platform immunity would only cover the removal of posts that fall within the narrowly specified categories of harm, enumerated in section 230(c)(2)(A).[122] To obtain this immunity, platforms would have to demonstrate an objectively reasonable belief that the content removed falls within one of those categories.[123] The OFVDA was introduced on September 8, 2020, and is currently before the Senate Committee on Commerce, Science, and Transportation.[124]

In 2019, Representative Louie Gohmert introduced the Biased Algorithm Deterrence Act to remove section 230 protection for platforms that "hinder" the display of user-generated content.[125] In the same year, Representative Paul Gosar proposed the Stop the Censorship Act to combat a perceived conservative bias in content moderation practices.[126] Under Senator Gosar's proposal, in order to retain section 230 immunity, social networks could only remove *illegal* content from their platforms. As a result, any platform whose content moderation practices involve removing legal, yet objectionable material, would lose its immunity.[127] Many of these efforts to amend sec-

119. Press Release, U.S. Senate Comm. on Com., Sci., & Transp., Wicker, Graham, Blackburn Introduce Bill to Modify Section 230 and Empower Consumers Online (Sept. 8, 2020), https://www.commerce.senate.gov/2020/9/wicker-graham-blackburn-introduce-bill-to-modify-section-230-and-empower-consumers-online [https://perma.cc/68GX-G4GE].

120. *Id.*

121. Online Freedom and Viewpoint Diversity Act, S. 4534, 116th Cong. (2020).

122. U.S. Senate Comm. on Com., Sci., & Transp., *supra* note 119.

123. *See id.*

124. S. 4534, 116th Cong. (2020).

125. Biased Algorithm Deterrence Act of 2019, H.R. 492, 116th Cong. (2019).

126. Stop the Censorship Act, H.R. 4027, 116th Cong. (2019); Press Release, Paul Gosar, D.D.S., Representing Ariz.'s 4th Dist., Gosar Introduces Stop the Censorship Act of 2020 (July 29, 2020), https://gosar.house.gov/news/documentsingle.aspx?DocumentID=3968#:~:text=Gooden.,and%20complements%20President%20Donald%20J [https://perma.cc/E6AM-47CB].

127. H.R. 4027, 116th Cong. (2019).

tion 230 are based on the false belief that section 230 somehow requires platform neutrality, which it plainly does not.[128]

2. Calls to Leverage Immunity in Exchange for Social Responsibility

Criticisms of section 230 are not limited to conservatives, though the arguments for revision of the law among liberals typically find their grounding in beliefs that the law provides too much immunity for social platforms that do too little to regulate harmful speech.[129] A chief concern for democratic lawmakers is the rampant disinformation that spread on social platforms amidst the 2016 presidential election, and that seemingly little has been done to address it since that time.[130]

It is true that the spread of false information on social platforms is unique compared to other forms of media. Unlike traditional news media that depend on fact-checking and verification processes, social platforms are built for instant content sharing, without verification.[131] Indeed, their business models *rely* on a lack of gatekeepers,[132] and "[t]he information-sharing environment is well suited to the spread of falsehoods."[133] President Joe Biden has been an outspoken critic of section 230 for this reason, arguing that immunity should be revoked for social platforms in light of the fact that Facebook and other large social platforms knowingly allowed falsehoods to propagate online in

128. *See* 141 CONG. REC. H8469–70 (daily ed. Aug. 5, 1996) (statement of Rep. Cox), https://www.congress.gov/congressional-record/1995/8/4/house-section/article/h8460-1 [https://perma.cc/QEX9-6TKT]; *see also* Emily Stewart, *Ron Wyden Wrote the Law That Built the Internet. He Still Stands by It—And Everything It's Brought with It.*, VOX (May 16, 2019, 9:50 AM), https://www.vox.com/recode/2019/5/16/18626779/ron-wyden-section-230-facebook-regulations-neutrality [https://perma.cc/8K8M-UAA9] (quoting Senator Ron Wyden, author of section 230, as saying "Section 230 is not about neutrality. Period. Full stop."); Weissman, *supra* note 95; Harmon, *supra* note 105.

129. Marguerite Reardon, *What's Section 230? The Social Media Law That's Clogging up the Stimulus Talks*, CNET (Dec. 30, 2020, 8:16 AM), https://www.cnet.com/news/whats-section-230-the-social-media-law-thats-clogging-up-the-stimulus-talks/ [HTTPS://PERMA.CC/EJ3N-DE5Q].

130. *See* Andrew Hutchinson, *Despite Everything, Facebook Remains a Prominent Facilitator of Election Misinformation*, SOCIAL MEDIA TODAY (Nov. 17, 2020), https://www.socialmediatoday.com/news/despite-everything-facebook-remains-a-prominent-facilitator-of-election-mi/589255/ [HTTPS://PERMA.CC/GB9J-HRBX].

131. JOE GREEK, SOCIAL NETWORK-POWERED INFORMATION SHARING 36 (2014).

132. Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 19–20 (2020); Marguerite Rigoglioso, *Online Social Networks Can Increase Ad Revenue by Stimulating Content*, STAN. BUS. (Nov. 1, 2011), https://www.gsb.stanford.edu/insights/online-social-networks-can-increase-ad-revenue-stimulating-content [https://perma.cc/DP2R-ZMV3] (explaining that social platforms have a financial incentive to encourage their users to post as much content as possible).

133. Robert Chesney & Danielle K. Citron, *Disinformation on Steroids: The Threat of Deep Fakes*, COUNCIL ON FOREIGN RELS. (Oct. 16, 2018), https://www.cfr.org/report/deep-fake-disinformation-steroids [https://perma.cc/V4AS-J53D].

the form of false political advertising.[134] President Biden's end game is focused on *more* content moderation geared toward reducing misinformation, in contrast to Mr. Trump's apparent position that content should be left entirely unmoderated.[135] Other democrats have voiced similar concerns.[136] House Speaker Nancy Pelosi threatened in 2019 that section 230 could be "in jeopardy" based on an argument that social platforms have not been "treating it with the respect that they should."[137] Senators Chris Coons and Mark Warner have also warned of imposing regulation on social platforms unless they meaningfully address issues concerning the spread of misinformation that threatens the democratic process.[138] The role of social platforms in spreading such misinformation and its impact on the 2016 presidential election is well-documented.[139]

Harms beyond falsehoods also spread on social platforms and include defamation, revenge pornography, hate speech, harassment directed towards marginalized groups, and more.[140] Scores of commentators and legislators have argued that the proliferation of this problematic speech is the result of section 230, and have called for its amendment or outright revocation.[141] Vice President Kamala Har-

134. Makena Kelly, *Joe Biden Doesn't Like Trump's Twitter Order, but Still Wants to Revoke Section 230*, THE VERGE (May 29, 2020, 1:50 PM), https://www.theverge.com/2020/5/29/21274812/joe-biden-donald-trump-twitter-facebook-section-230-moderation-revoke [https://perma.cc/8MZY-DZLY]; Cristiano Lima, *Biden: Tech's Liability Shield 'Should Be Revoked' Immediately*, POLITICO (Jan. 17, 2020, 10:56 AM), https://www.politico.com/news/2020/01/17/joe-biden-tech-liability-shield-revoked-facebook-100443 [https://perma.cc/P9KW-39X9].

135. Lima, *supra* note 134.

136. Todd Shields & Ben Brody, *Washington's Knives are Out for Big Tech's Social Media Shield*, BLOOMBERG (Aug. 11, 2020, 3:00 AM), https://www.bloomberg.com/news/articles/2020-08-11/section-230-is-hated-by-both-democrats-and-republicans-for-different-reasons [https://perma.cc/BA2A-H3H4].

137. Taylor Hatmaker, *Nancy Pelosi Warns Tech Companies That Section 230 Is 'In Jeopardy'*, TECHCRUNCH (Apr. 12, 2019, 2:35 PM), https://techcrunch.com/2019/04/12/nancy-pelosi-section-230/ [https://perma.cc/98V2-73H6].

138. Bloomberg, *Senators Threaten to Regulate Facebook Unless It Makes Fixes*, FORTUNE (Nov. 17, 2018, 7:24 AM), https://fortune.com/2018/11/17/facebook-regulation-privacy/ [https://perma.cc/BY2Y-5F64]; Mark R. Warner, *Potential Policy Proposals for Regulation of Social Media and Technology Firms*, SENATE.GOV (July 30, 2018), https://www.warner.senate.gov/public/index.cfm/2018/7/read-the-warner-social-media-white-paper [https://perma.cc/CSH2-DW6X].

139. Hasen, *supra* note 9, at 205–07; Morgan Chalfant, *Researchers Say Fake News Had 'Substantial Impact' on 2016 Election*, THE HILL (Apr. 3, 2018, 2:21 PM), https://thehill.com/policy/cybersecurity/381449-researchers-say-fake-news-had-substantial-impact-on-2016-election [https://perma.cc/MRU2-U29C]; Alexis C. Madrigal, *What Facebook Did to American Democracy*, THE ATLANTIC (Oct. 12, 2017), https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/ [https://perma.cc/HPS3-79KR].

140. Paul M. Barrett, Opinion, *Why the Most Controversial US Internet Law Is Worth Saving*, MIT TECH. REV. (Sept. 9, 2020), https://www.technologyreview.com/2020/09/09/1008223/section-230-internet-law-policy-social-media-misinformation/ [https://perma.cc/CLB3-YZJV].

141. *Id.*

ris, for example, has made efforts to revise or repeal section 230 for years, since she was the attorney general for California, based primarily upon concerns about the use of online spaces by child sex traffickers.[142] Professors Danielle Citron and Benjamin Wittes argue that section 230 immunity is too sweeping given that it protects platforms that knowingly create avenues for sexual predators to connect with victims.[143] Professor Citron has suggested a balanced approach that would recognize cyber rights while allowing for protections of section 230,[144] and Professor Vanessa Browne-Barbour argues that courts should adopt a narrower interpretation of section 230 to provide a defamation remedy.[145] Some commentators have recommended that section 230 be amended to more closely resemble the Digital Millennium Copyright Act's notice-and-takedown policy by making civil immunity contingent upon the site taking down offensive content once brought to its attention.[146] Others have suggested that section 230 be amended to specifically exclude social platforms from civil immunity based on the rationale that they are not merely passive service providers, but instead play a role in selecting what information is spread on their platforms.[147]

A bipartisan group of legislators led by Senator Lindsey Graham introduced the EARN IT Act of 2020 ("EARN IT"), which is an effort to hold social platforms accountable for the child exploitation that exists on their platforms.[148] This act would operate similarly to FOSTA by denying social networks immunity for causes of action arising from user content that sexually exploits minors.[149] EARN IT would also create a panel to advise social platforms on best practices to curb child exploitation on their sites.[150] EARN IT's critics have expressed concern that the proposed legislation is actually a subtle attack on encryption.[151] This is because social networks may be incentivized to ban encryption to better monitor all content on their

142. Elizabeth Nolan Brown, *Section 230 Is the Internet's First Amendment. Now Both Republicans and Democrats Want to Take It Away.*, REASON (July 29, 2019, 8:01 AM), https://reason.com/2019/07/29/section-230-is-the-internets-first-amendment-now-both-republicans-and-democrats-want-to-take-it-away/ [https://perma.cc/WF96-LD5P].

143. Citron & Wittes, *supra* note 31.

144. Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 117 (2009).

145. Vanessa S. Browne-Barbour, *Losing Their License to Libel: Revisiting § 230 Immunity*, 30 BERKELEY TECH. L.J. 1505, 1505 (2015).

146. Andrew P. Bolson, *Flawed but Fixable: Section 230 of the Communications Decency Act at 20*, 42 RUTGERS COMPUT. & TECH. L.J. 1, 14–15 (2016).

147. Nicole Phe, *Social Media Terror: Reevaluating Intermediary Liability Under the Communications Decency Act*, 51 SUFFOLK U. L. REV. 99, 127 (2018).

148. Bedell & Major, *supra* note 108.

149. *Id.*

150. *Id.*

151. Lily Hay Newman, *The EARN IT Act is a Sneak Attack on Encryption*, WIRED (Mar. 5, 2020, 8:22 PM), https://www.wired.com/story/earn-it-act-sneak-attack-on-encryption/ [https://perma.cc/92Q7-EYF9].

platforms and thereby avoid liability.[152] Nevertheless, the Senate Judiciary Committee approved EARN IT.[153] One June 24, 2020, Senator Brian Schatz introduced another effort, the Platform Accountability and Consumer Transparency Act.[154] It has since been referred to the Committee on Commerce, Science, and Transportation.[155] The proposal is intended to reduce illegal content on social media networks and simultaneously promote consistency and transparency within content moderation practices.[156]

The point is simply this: there are as many ideas to redefine, limit, or otherwise amend section 230 as there are speakers on the subject.[157] This Article is neither an attempt to unpack those ideas nor compile an exhaustive list of them. Instead, it is an attempt to recognize that section 230 has had myriad critics as well as supporters, discussed in Part III.B, complicating efforts to achieve consensus on what, if any, reform should look like.

3. The Nexus Between Content Regulation and Section 230 Reform

Lawmakers and scholars are calling for revision of section 230 to achieve changes to the type of content social media companies allow on their platforms.[158] On one hand, this makes perfect sense because enacting a new law to restrict certain content on social platforms would suffer from certain constitutional challenges.[159] Section 230, on the other hand, is a congressional grant that offers valuable immunity for platforms.[160] Predicating section 230 protection upon moderation of content in a certain manner allows for a framework for regulating content while avoiding constitutional problems.

---

152. *Id.*

153. Makena Kelly, *A Weakened Version of the EARN IT Act Advances Out of Committee*, THE VERGE (July 2, 2020, 12:44 PM), https://www.theverge.com/2020/7/2/21311464/earn-it-act-section-230-child-abuse-imagery-facebook-youtube-lindsey-graham [https://perma.cc/FBV9-DDVH].

154. S. 4066, 116th Cong. (2020).

155. *Id.*

156. *Schatz, Thune Introduce New Legislation to Update Section 230, Strengthen Rules, Transparency on Online Content Moderation, Hold Internet Companies Accountable for Moderation Practices*, BRIAN SCHATZ U.S. SENATOR FOR HAW. (June 24, 2020), https://www.schatz.senate.gov/press-releases/schatz-thune-introduce-new-legislation-to-update-section-230-strengthen-rules-transparency-on-online-content-moderation-hold-internet-companies-accountable-for-moderation-practices [https://perma.cc/DG4Q-3BWV].

157. *See* Klonick, *supra* note 9, at 1613–14 (collecting a thorough list of legal scholars' positions).

158. *See supra* Part II.B (discussing how numerous legislators and commentators suggested amending section 230 to limit platform immunity).

159. *See supra* Part II.B (discussing how numerous legislators and commentators suggested amending section 230 to limit platform immunity).

160. Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 427, 433–34 (2009).

A threshold problem is that efforts to reform section 230 to induce particular content moderation practices will necessarily make distinctions about speech on the basis of its content. For example, if revisions to section 230 are predicated on reducing false information, the government is making what would otherwise be an impermissible speech distinction: false speech is often protected by the First Amendment.[161] Furthermore, who would determine whether the platform made the correct decision regarding what constituted false information? Would this role fall to someone at a government agency like the FTC,[162] a judge in an action where the social platform has pled a section 230 defense and the plaintiff claims the defendant waived immunity for failure to comply with the law, or another government representative? None of these options are ideal because the question of falsity is left either to a government official—exceptionally troubling from a First Amendment perspective—or creates the expense of litigating the truth or falsity of a statement on a motion to dismiss.[163] This same dilemma would be true for other types of problematic speech identified by lawmakers and commentators as reasons to revisit section 230 immunity.

This is particularly problematic given the importance social platforms have in public discussion and debate. Posts on controversial topics or those with alternative viewpoints would be especially vulnerable to removal, which has the practical effect of deputizing online companies to censor speech that the government would never be allowed to touch because of the First Amendment.[164] The result, in short, is de facto government regulation of online speech. This is the chief reason why amending section 230 to force platforms into better accountability is so problematic: the government is able to make an end run around the First Amendment's restrictions on creating content-based regulations.

Another problem with this type of reform is that it is built on the faulty assumption that conditioning platform immunity on content moderation will achieve a net reduction of online speech harms. In all likelihood, speech harms will continue to exist even in a world without section 230 protection. Removing this immunity (or threatening to remove it by making its protections contingent upon satisfaction of certain requirements) may indeed create incentives for platforms to

161. Brown & Peters, *supra* note 4, at 533.

162. The May 28, 2020 executive order on Preventing Online Censorship purports to designate the FTC as the decisionmaker of whether complaints allege violations of law that implicate the policies set forth by the order. Exec. Order No. 13,925, 85 Fed. Reg. 34,079, 34,082 (May 28, 2020).

163. Cass. R. Sustein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 387, 398–400 (2020), https://jolt.law.harvard.edu/assets/articlePDFs/v33/33HarvJLTech387.pdf [HTTPS://PERMA.CC/W896-BCAT].

164. Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1364–66 (2018).

minimize speech harms, but this will come at a steep cost: over-removal of speech.[165] The bottom line is that penalizing platforms by removing section 230 protection does not induce platforms to create a safer space—it induces them to minimize risk.

Faced with the practical impossibility of removing all harmful speech, platforms would have to draw lines to remove user speech that is likely to lead to liability for the platform, while leaving in place speech that will not. Current content moderation practices favor automated decision making and rely less on human interpretation.[166] This creates challenges in differentiating between speech that creates a potential legal liability, and that speech which would be protected. Indeed, "[a] key concern in the deployment of automated moderation technologies in the context of copyright is systematic overblocking."[167] The reason is simple: many moderation decisions require nuanced legal analysis; such as determining whether a negative statement about someone rose to the level of actionable defamation, or whether posted information about a sexual assault could be a disclosure of another's private fact, or whether a deepfake was legally problematic or a protected parody, and so on.[168] As more platforms move from human to algorithmic moderation,[169] they are less likely to get these tough context-based decisions—such as differentiating between a woman breastfeeding and posing for a topless photo—correct,[170] and at the same time would be faced with an increased risk of liability for getting those decisions wrong. "The clearest problem is that language is incredibly complicated, personal and context dependent: even words that are widely accepted to be slurs may be used by members of a group to reclaim certain terms."[171] This is why "[c]ontent moderation at scale is impossible to perform perfectly—platforms have to make millions of decisions a day and cannot get it right in every instance. Because error is inevitable, content moderation system

165. Jeff Kosseff, *Defending Section 230: The Value of Intermediary Liability*, 15 J. TECH. L. & POL'Y 123, 151–52 (2010).

166. Klonick, *supra* note 9, at 1636.

167. Gorwa, Binns & Katzenbach, *supra* note 27, at 7–8.

168. *Id.*

169. *See* Sarah T. Roberts, *The Great A.I. Beta Test*, SLATE (Apr. 8, 2020, 12:53 PM), https://slate.com/technology/2020/04/coronavirus-facebook-content-moderation-automated.html [https://perma.cc/V89M-C9RZ] (describing how policy exemptions will not be understood by algorithms, such as when "material that would look, to a machine, like excessive blood and gore but, in fact, was the video of an unlawful attack on civilians in a conflict zone").

170. *See id.* (noting that "[d]espite their technological sophistication, such automated tools fall far short of a human's discernment"); Gorwa, Binns & Katzenbach, *supra* note 27, at 8 (describing that "[w]hile Content ID and other systems may improve from a technical standpoint, enhancing their ability to create quality fingerprints and then accurately detect those fingerprints, it does not necessarily mean that they become more adept at evaluating *actual* copyright infringement" (emphasis in original)).

171. Gorwa, Binns & Katzenbach, *supra* note 27, at 10.

design requires choosing which kinds of errors the system will err on the side of making."[172] With a threat of liability looming—in spades, because removing section 230 opens the floodgates to plaintiffs hungry for a deep-pocketed defendant—this framework incentivizes platforms to remove all speech that could be interpreted near that line.[173] This increased legal pressure on social platforms almost certainly would result in "overly aggressive, unaccountable self-policing, leading to arbitrary and unnecessary restrictions on online behavior."[174]

Indeed, platforms have acknowledged that over-censorship will occur because the automated moderation systems will find more "'false positives,' meaning that content will be removed that should remain up."[175] The history of platforms' compliance with the DMCA by over-removal illustrates this point—the threat of secondary liability induces service providers to comply with the notice-and-takedown provisions, even when the notice is questionable or flawed.[176] There are also significant abuses of the takedown provision of the DMCA designed to silence speech. As Wendy Seltzer notes, the promise of rapid takedown compounds the problem and

> creates an incentive for copyright claimants to file dubious takedown claims. The mechanism is cheap for the claimant, more expensive for the respondent, and if the process stops after the claim stage (as it often does) the complained-of material remains offline. And unless the complaint is so groundless that it can give rise to a lawsuit against the complainant, a non-infringing poster has no legal or practical recourse against bogus claims.[177]

None of this is to say that the goals of reducing or removing harmful speech online are not important—indeed they are. And it may well be true that the only means to achieve an Internet with fewer speech harms is through an imperfect system of content moderation. But overregulation at the request of—or inducement by—the government is inherently more problematic than a platform's own decision to over-remove content. Any approach to amend section 230 to achieve

172. Evelyn Douek, *COVID-19 and Social Media Content Moderation*, LAWFARE (Mar. 25, 2020, 1:10 PM), https://www.lawfareblog.com/covid-19-and-social-media-content-moderation [https://perma.cc/T93W-2UC2].

173. Kosseff, *supra* note 165. Even where the legal framework requires "good faith" efforts to remove harmful content, the platform is still incentivized to over-remove potentially objectionable content so that there is little argument that it did not act in consistent good faith. *Id.* at 131–32.

174. Milton L. Mueller, Comment, *Hyper-Transparency and Social Control: Social Media as Magnets for Regulation*, 39 TELECOMMS. POL'Y 804, 809 (2015); Roberts, *supra* note 169 ("The overly broad bluntness of these tools is less of a mistake and more of an infringement on the right to create, access, and circulate information.").

175. Douek, *supra* note 172.

176. Wendy Seltzer, *Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of the DMCA on the First Amendment*, 24 HARV. J.L. & TECH. 171, 177 (2010).

177. *Id.* at 178.

content regulation goals would likely induce the removal of too much speech—instead of encouraging platforms to remove as much bad speech as possible—while recognizing that getting it right all of the time is impractical.

III. Finding the Best Way Forward

While achieving government-driven content regulation through section 230 reform is problematic, reducing speech harms on social platforms is an important goal, and one that social media companies should be encouraged to devote their resources toward. The challenge is providing the appropriate incentives to social platforms that will hold them accountable while considering countervailing First Amendment interests.

A key concern related to any content regulation goal or requirement for social platforms is an understanding of what is technologically possible and how social media companies are using that technology in their current content moderation practices. This is a critically important and ever-evolving area that, while generally beyond the scope of this Article, warrants a brief discussion.

A. *Methods of Content Moderation*

The major social networks have taken a mixed approach to content moderation, using both algorithms and humans to remove harmful content from their platforms. Sites such as Facebook, Twitter, and YouTube use algorithms to detect content suspected of violating their community standards.[178] Algorithms then flag the questionable posts and refer them to human content moderators for evaluation.[179]

Algorithms are an important tool in moderating social platforms, but they have limitations. This is particularly true when it comes to identifying whether speech expressing social or political views is permissible. This limitation has meant that algorithms have struggled to accurately identify—and remove—hate speech.[180] Algorithms have

178. *How Automated Tools Are Used in the Content Moderation Process*, New Am., https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/how-automated-tools-are-used-in-the-content-moderation-process/ [https://perma.cc/LF7Y-5GVF].

179. Sissi Cao, *Facebook's AI Chief Explains How Algorithms Are Policing Content—And Whether It Works*, Observer (Dec. 6, 2019, 7:15 AM), https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/ [https://perma.cc/6GGC-7XDH]; *How Automated Tools Are Used in the Content Moderation Process*, *supra* note 178.

180. *See, e.g.*, Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech, but Not Black Children*, ProPublica (June 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/S5KG-C54B] (noting that Facebook algorithms classified a post calling for the killing of all "radicalized" Muslims as protected political opinion, rather than removing it as a form of hate speech).

also repeatedly failed to intercept violent content before it appears on platforms.[181] Numerous abusive posts slide through the algorithmic cracks on a daily basis, leaving users on social platforms to report disturbing content after it has already circulated.[182] For example, Facebook's algorithms infamously allowed a live video of the Christchurch massacre to stream across the platform in 2019.[183] When Facebook finally terminated the livestream—in response to numerous user complaints—it had already run for seventeen minutes.[184]

Just as algorithms fail to detect offensive posts, they frequently censor appropriate content. For instance, after a community newspaper in Texas posted excerpts from the Declaration of Independence, Facebook's hate speech algorithm flagged and automatically removed the post for violating the platform's hate speech standards.[185] The reason was likely a line in the document about "merciless Indian savages," though Facebook declined to confirm that was the case.[186] The same is true when it comes to images of nudity. Although algorithms have been relatively successful at identifying nudity, they have proven incapable of discerning the harmless from the harmful. In addition to removing child pornography, algorithms have erroneously removed images of classic art and women breastfeeding.[187]

These errors establish a major shortcoming of algorithmic content moderation: its inability to understand context. While algorithms may be useful tools for flagging purposes, they are not a substitute for human content moderators. Unlike human content moderators, algo-

181. *See* James Vincent, *AI Won't Relieve the Misery of Facebook's Human Moderators*, THE VERGE (Feb. 27, 2019, 12:41 PM), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms [https://perma.cc/HN5C-3LBE].

182. Kalev Leetaru, *Facebook's Failed AI Showcases the Dangers of Technologists Running the World*, FORBES (Mar. 22, 2019, 11:56 AM), https://www.forbes.com/sites/kalevleetaru/2019/03/22/facebooks-failed-ai-showcases-the-dangers-of-technologists-running-the-world/#5902277639cb [https://perma.cc/54ZX-U2UK].

183. *Id.*

184. Donie O'Sullivan, *Facebook Says It's Policing Its Platform, but It Didn't Catch a Livestream of a Massacre. Why?*, CNN BUS., https://www.cnn.com/2019/03/15/tech/facebook-new-zealand-content-moderation/index.html (Mar. 15, 2019, 2:21 PM) [https://perma.cc/V8Q8-3HZZ].

185. Sam Wolfson, *Facebook Labels Declaration of Independence As 'Hate Speech'*, THE GUARDIAN (July 5, 2018, 1:10 PM), https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech [https://perma.cc/RK6G-G3DJ].

186. *Id.*

187. Stephanie Linning, *'It Is The Most Beautiful and Natural Thing': Facebook Sparks User Backlash When It Removes Photo of Topless Mother Breastfeeding Her Baby*, MAIL ONLINE (Jun. 29, 2017, 1:55 PM), https://www.dailymail.co.uk/femail/article-4650418/Facebook-user-ordered-remove-photo-breastfeeding-mum.html [https://perma.cc/2K3H-FYAQ].

rithms take posts at face value.[188] The algorithms consider only what is being said, paying little regard to the post's purpose or what it *actually* communicates to the platform's audience. For this reason, algorithms have proven consistently incapable of understanding linguistic nuances such as humor and sarcasm.[189]

Because algorithms are not perfect, there is often a call for more human content moderation to ensure platforms remove harmful content but do not wrongfully censor harmless—or even socially beneficial—content. Although humans are often in a better position than a machine to determine whether speech is prohibited by community standards, this too is an imperfect solution. Human moderators also struggle with difficult decisions and apply community standards inconsistently—a product of vague guidelines, broad discretion, and their own subjective biases.[190] In addition, social platforms rely on humans to make decisions about speech that often straddle the line between permissible and impermissible. Unless they are trained attorneys, human moderators continuously struggle with decisions over what is or is not legally harmful speech.[191]

In addition, when it comes to certain types of harmful content such as violence and nudity, viewing hours of disturbing content takes a heavy psychological toll on human moderators.[192] In fact, some of the major social networks are now mandating that their employees sign forms, acknowledging that their work as content moderators could cause PTSD.[193] As a result, platforms increasingly rely on algorithms to remove this content without human review.[194] Despite their individual shortcomings, the joint efforts of algorithmic and human con-

188. Ben Dickson, *Why AI Is Terrible at Content Moderation*, MAGZTER (Sept. 2019), https://www.magzter.com/article/Science/PC-Magazine/Why-AI-Is-Terrible-At-Content-Moderation [https://perma.cc/ZF8K-544J].

189. *See* Vincent, *supra* note 181.

190. *See* Jillian C. York & Corynne McSherry, *Content Moderation Is Broken. Let Us Count the Ways.*, EFF (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways [https://perma.cc/8ZNQ-XYSH]; Daisy Soderberg-Rivkin, Opinion, *When It Comes to Content Moderation, We've Been Focusing on the Wrong Type of Bias*, MORNING CONSULT (Dec. 5, 2019, 5:00 AM), https://morningconsult.com/opinions/when-it-comes-to-content-moderation-weve-been-focusing-on-the-wrong-type-of-bias/ [https://perma.cc/Z3DB-6AXQ].

191. *See* Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1361–62 (2018).

192. Alex Hern, *Revealed: Catastrophic Effects of Working As a Facebook Moderator*, THE GUARDIAN (Sep. 17, 2019), https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator [https://perma.cc/X63Y-AEBM].

193. *Facebook and YouTube Moderators Sign PTSD Disclosure*, BBC NEWS (Jan. 25, 2020), https://www.bbc.com/news/technology-51245616 [https://perma.cc/CMQ2-4BSY].

194. Vincent, *supra* note 181.

tent moderators present the most effective approach to fighting online abuses in light of the current technology available.[195]

B. *Models for Content Regulation*

Unlike the *methods* for moderating content, which rely on a combination of algorithmic and human decision-making, there are three different models for making determinations about how content is moderated. Under the current model, platforms are responsible for devising their own content moderation systems. It would be inaccurate to describe this as a "zero-regulation" framework, given that platforms engage in a fair amount of content moderation, as discussed above, but this system of self-regulation allows platforms incredible discretion over moderating content. The way social platforms have exercised this discretion has led to calls for *increased* governance, particularly for policies related to the spread of misinformation and hate speech. As discussed in Part II.B, several lawmakers and commentators have suggested reforming section 230 to achieve these policy goals, and others—including Facebook's co-founder and CEO Mark Zuckerberg—have called for direct government regulation over social media.[196] There also exists a third regulatory option to achieve the reduction of speech harms online—an industry-wide self-governance model.

Thus, the three distinct models are: (1) the current model of self-regulation, where platforms devise individual schemes for content moderation; (2) the government regulation model; and (3) the industry-wide self-governance model. This Part will offer a high-level overview of each model.

1. Self-Regulation: Content Moderation That Is Too Small

Despite the fact that social platforms are not legally required to engage in content moderation, there are important normative and economic considerations that create incentives to self-regulate. Although social platforms remain free from *government* regulation, self-regulation is inherently a model of private regulation.[197] This model would likely gain support if there existed public certainty that social platforms did all in their power to remove harmful speech, and if there was a consensus regarding what constitutes harmful speech. The chal-

195. *See* Dickson, *supra* note 188; *see also* Kalev Leetaru, *Why We Still Need Human Moderators in an AI-Powered World*, FORBES (Sept. 8, 2018, 9:42 AM), https://www.forbes.com/sites/kalevleetaru/2018/09/08/why-we-still-need-human-moderators-in-an-ai-powered-world/#7df900541412 [https://perma.cc/P98W-PX5D].

196. Zuckerberg claimed government regulation is necessary "in four areas: harmful content, election integrity, privacy and data portability." Zuckerberg, *supra* note 30.

197. *See* Klonick, *supra* note 9, at 1662 (referring to private social platforms as systems of governance).

lenge is that neither condition exists: social platforms moderate content with little transparency,[198] and there is no clear consensus over what speech meets a threshold level of *harmful* to warrant removal.[199] The first challenge could be easily solved, but at a cost social platforms have to date been unwilling to bear.

This lack of transparency is problematic because there is no way to hold a platform accountable when the public does not know the decisions guiding its moderation practices. "Every time someone uses search or social media services, [they are] relying on a secret and proprietary algorithm tuned to maximize something—usually user engagement with the service. Transparency and accountability are largely absent."[200] Instead, the public is told, "Trust us, the engineers are working on it,"[201] which does little to inspire confidence given that social platforms constantly struggle with moderation decisions and speech harms still abound online. The lack of transparency breeds an additional problem because when it is not clear what social platforms are actually doing, there is no meaningful way to engage in a discussion about whether platforms are doing *enough*.

Another challenge with the self-regulation framework is that the business model for social platforms—which prioritizes engagement of as many users as possible—also favors controversy.[202] Any content that evokes emotion—good or bad—stimulates user engagement.[203] As a result, even harmful posts inciting negative emotions such as

198. Rebecca MacKinnon, *Facebook Is Part of an Industry-Wide Problem: Lack of Transparency About Policies Affecting Users' Online Rights*, BUS. & HUM. RTS. RESOURCE CENTRE (April 25, 2019), https://www.business-humanrights.org/en/facebook-is-part-of-an-industry-wide-problem-lack-of-transparency-about-policies-affecting-users%E2%80%99-online-rights [https://perma.cc/V358-R5XC]; Prerna Juneja, Deepika Rama Subramanian & Tanushree Mitra, *Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices*, 4 PACM HUM.-COMPUT. INTERACTION, Jan. 2020, at 3 (demonstrating an in-depth study conducted on Redditt's content moderation with results revealing a lack of transparency in moderation practices); David Siegel & Rob Reich, Opinion, *It's Not Too Late for Social Media to Regulate Itself*, WIRED (Feb. 7, 2019, 9:00 AM), https://www.wired.com/story/its-not-too-late-for-social-media-to-regulate-itself/ [https://perma.cc/AC5Q-V2BZ].

199. *See generally* Claire Wardle, *Challenges of Content Moderation: Define "Harmful Content"*, INSTITUT MONTAIGNE (June 27, 2019), https://www.institutmontaigne.org/en/blog/challenges-content-moderation-define-harmful-content [https://perma.cc/FR4B-5HFM]. This is the ultimate hurdle to achieving a system of workable content moderation. The threshold of what is harmful enough to warrant removal is a question of line-drawing that reasonable people will differ on. Not to mention the challenge of divorcing political biases from such moderation decisions (considering the adage "one man's freedom fighter on film is another man's terrorist").

200. *Id.*

201. Siegel & Reich, *supra* note 198.

202. Brown, *supra* note 132, at 21.

203. Martin Jones, *Emotional Engagement Is the Key to Viral Content Marketing*, COX BLUE, https://www.coxblue.com/emotional-engagement-is-the-key-to-viral-content-marketing/ [https://perma.cc/54DH-DWLM].

rage, fear, or disdain generate numerous shares on social media.[204] Social networks are not only aware of this correlation, they actively exploit it to maximize their profits.[205] For example, when Facebook learned how its algorithms contributed to polarizing its users, it eliminated efforts to make the platform less divisive, because controversy equals engagement.[206] Contentious posts will necessarily engage users, which may impact algorithmic decisions regarding display and promotion of that post.[207]

This is not to say that social platforms have not made real strides toward reducing harmful content; however, it is impossible to ignore their prioritization of economic objectives.[208] Social platforms, after all, are profit-oriented organizations that rely on maximizing user engagement to attract advertisers.[209] But not all decisions are entirely profit-driven. While certain decisions regarding the moderation of harmful content may be viewed as a component of a broader Corporate Social Responsibility ("CSR"), it would be more accurate to view many decisions through a hybrid CSR-economic lens.[210] Where market forces and CSR align, the self-regulation model *works* to remove harmful speech. Actions (or inactions) that alienate users could drive down potential advertising revenue, so social platforms have "developed an intricate system to both take down content their users don't want to see and keep up as much content as possible."[211] Many platforms have long emphasized their free speech values, but have had to weigh these "against competing principles of user safety, harm to users, public relations concerns . . . and the revenue implications of certain content for advertisers."[212]

The current decentralized nature of self-governance has created opportunities for each platform to make the best content decisions for its

204. *Id.*

205. Jeff Horwitz & Deepa Seetharaman, *Facebook Executives Shut Down Efforts to Make the Site Less Divisive*, WALL ST. J. (May 26, 2020, 11:38 AM), https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?campaign_id=9&emc=edit_nn_20200527&instance_id=18835&nl=the-morning&regi_id=72867838&segment_id=29269&te=1&user_id=dd9f065f08eef4679935efacda7d37b2 [https://perma.cc/N3PE-H3H6].

206. *Id.*

207. *Id.*

208. Klonick, *supra* note 9, at 1627.

209. Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939, 953 (2020) ("Increasing user engagement is financially lucrative for online platforms: as users spend more time and attention on their sites, platforms can collect ever more behavioral data, improve their targeted advertising and engagement capabilities, and grow their advertising revenue.").

210. Klonick, *supra* note 9, at 1627 ("Though corporate responsibility is a noble aim, the primary reason companies take down obscene and violent material is the threat that allowing such material poses to potential profits based in advertising revenue.").

211. *Id.* at 1664.

212. *Id.* at 1626.

particular users. This enables social media sites to take unique approaches to striking their desired balance of "figur[ing] out new approaches or rules that would still satisfy concerned users and encourage them to connect and interact on the platform."[213] The model has allowed platforms to individually curate content in ways that make the most sense for their users:

> This decentralization allows some sites to focus on providing an experience that feels safe, or entertaining, or suitable for kids, while others aim to foster debate, or create an objective encyclopedia, or maintain an archive of videos documenting war crimes. Each of these is a distinct and laudable goal, but each requires different content standards and moderation practices.[214]

Such goals are possible given the lack of government control. This inherent flexibility has also meant that some platforms take more aggressive approaches against harmful speech while others do little.[215]

Unsurprisingly, social platforms tend to invest resources in removing harmful speech when doing so offers a clear economic benefit.[216] For example, up until 2015, terrorist organizations openly exploited social platforms like Twitter, YouTube, and Facebook to organize, recruit, fundraise, and inspire violence.[217] Initially, these platforms were slow to react, but their receiving threats of legal action from around the world tipped the needle.[218] Beginning in 2015, a series of lawsuits emerged against social platforms, including Twitter, Facebook, and YouTube, alleging that by allowing terrorist groups to use their platforms, the social media companies were providing material support to terrorists.[219] Although section 230 immunized platforms against these claims, the lawsuits both shined a spotlight on these abuses and increased public calls for meaningful change.[220]

---

213. *Id.*

214. Emma Llanso, *Platforms Want Centralized Censorship. That Should Scare You.*, WIRED (Apr. 18, 2019, 9:00 AM), https://www.wired.com/story/platforms-centralized-censorship/ [https://perma.cc/97RQ-ZZSA].

215. Brown, *supra* note 132, at 10.

216. Klonick, *supra* note 9, at 1627 (explaining how companies remove harmful content when refraining from doing so could adversely affect profits).

217. Brown, *supra* note 132, at 7.

218. *Id.* at 9, 11–12.

219. Crosby v. Twitter, Inc., 921 F.3d 617, 619 (6th Cir. 2019); Force v. Facebook, Inc., 934 F.3d 53, 68–69 (2d Cir. 2019); Taamneh v. Twitter, Inc., 343 F. Supp. 3d 904, 908 (N.D. Cal. 2018); Copeland v. Twitter, Inc., 352 F. Supp. 3d 965, 968 (N.D. Cal. 2018); Clayborn v. Twitter, Inc., No. 17-cv-06894-LB, 2018 WL 6839754, at *1 (N.D. Cal. Dec. 31, 2018); Cohen v. Facebook, Inc., 252 F. Supp. 3d 140, 147 (E.D.N.Y. 2017); Gonzalez v. Google, Inc., 282 F. Supp. 3d 1150, 1155 (N.D. Cal. 2017); Pennie v. Twitter, Inc., 281 F. Supp. 3d 874, 879 (N.D. Cal. 2017); Cain v. Twitter, Inc., 17 Civ. 122 (PAC), 2017 WL 1489220, at *1 (S.D.N.Y. Apr. 25, 2017); Fields v. Twitter, Inc., 217 F. Supp. 3d 1116, 1119 (N.D. Cal. 2016).

220. Klonick, *supra* note 9, at 1604; *see* Julia Greenberg, *Twitter Wants You to Know That It Is Fighting Terrorists*, WIRED (Feb. 5, 2016, 3:18 PM), https://www.wired.com/2016/02/twitter-wants-you-to-know-that-it-is-fighting-terrorists/ [https://perma.cc/2JU4-UE5L].

Twitter, reluctant to be seen as a tool of the government, initially took a completely hands-off approach, refusing to remove even those users requested by the U.S. government on the basis of their identification and designation as terrorists.[221] This began to change when governments around the world unrestrained by the First Amendment, particularly the European Union, applied legal pressure on platforms to do more to remove terrorist speech.[222] This was likely Twitter's impetus in 2016, when it changed course and began to dramatically increase the number of accounts it suspended for promoting terrorism.[223] Facebook also began actively policing terrorists' abuse of its service by removing millions of posts.[224] YouTube, arguably the top recruitment platform for terrorist networks, which had initially struggled to respond to the dissemination of terrorist content, eventually developed a more successful strategy for fighting the spread of its propaganda.[225] Eventually, these platforms, along with Microsoft, engaged in a collaborative effort to reduce extreme and egregious terrorist content online.[226]

These efforts to stem terrorist organizations' abuse of social platforms can be viewed as advancing both CSR and economic objectives: by committing resources to addressing a global challenge, they demonstrate corporate accountability and at the same time send a signal to users that they are working to maintain a safe online space.[227] But they also achieve an important goal of mitigating the threat of costly governmental sanctions.[228]

That there is not always perfect alignment between economic and CSR goals is a significant drawback to a model of self-regulation.[229]

221. Brown, *supra* note 132, at 10–11.

222. *See* Michal Lavi, *Do Platforms Kill?*, 43 HARV. J.L. & PUB. POL'Y 477, 507–09 (2020).

223. Greenberg, *supra* note 220; Queenie Wong, *Twitter Cracks Down on Accounts Promoting Terrorism*, MERCURY NEWS (Sept. 20, 2017, 3:51 AM), https://www.mercurynews.com/2017/09/19/twitter-cracks-down-on-accounts-promoting-terrorism/ [https://perma.cc/3BD3-9MKP].

224. *Id.*; Tony Romm & Elizabeth Dwoskin, *Facebook Says It Removed a Flood of Hate Speech, Terrorist Propaganda and Fake Accounts from Its Site*, WASH. POST (Nov. 15, 2018, 4:58 PM), https://www.washingtonpost.com/technology/2018/11/15/facebook-says-it-removed-flood-hate-speech-terrorist-propaganda-fake-accounts-its-site/ [https://perma.cc/7WXV-EXDF].

225. Rita Katz, *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.*, WIRED (Oct. 28, 2018, 8:00 AM), https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really/ [https://perma.cc/FMM6-9ZYP].

226. GLOBAL INTERNET FORUM TO COUNTER TERRORISM, https://www.gifct.org/ [https://perma.cc/QT4X-3R9G].

227. *See* Klonick, *supra* note 9, at 1626–27.

228. *See* United States v. Alvarez, 567 U.S. 709, 746 (Alito, J., dissenting) (2012) ("The constitutional guarantees of the First Amendment can tolerate sanctions against *calculated* falsehood without significant impairment of their essential function." (quoting *Garrison v. Louisiana*, 379 U.S. 64, 75 (1964)) (internal quotations omitted) (emphasis in original)).

229. *See* Siegel & Reich, *supra* note 198.

"History teaches us that unregulated marketplaces can produce a race to the bottom, externalizing harms while socializing these costs and privatizing the financial gains."[230] In this case, the negative externalities are borne not by the social platforms but by their users.[231] When social platforms promote division, false information, hate speech, or other harmful speech, their users bear the cost.[232] This is why a framework that leaves content moderation entirely up to social platforms is inherently problematic. Without any accountability or oversight, it offers no framework at all. The temptation to prioritize profits at the expense of all else is simply too great, and this is part of the driving call to lawmakers for legislative change.[233]

2. Government Regulation: Content Moderation That Goes Too Big

Just as there are a variety of models of private regulation, so too are there a variety of options for government-led content regulation. Lawmakers have already tried to advance frameworks where oversight could come from the FTC or FCC[234] and section 230 reform,[235] although no meaningful regulation yet exists. Without a clear proposed regulatory framework it is impossible to determine whether efforts to regulate would survive the requisite constitutional challenges.[236] For example, analyzing a law that would *restrict a platform's ability to remove* content connected to certain ideological viewpoints would be fundamentally different than an analysis *requiring a platform to take down* posts on the basis of their content. Despite the fact that it is inherently difficult to speculate about a generic model for government regulation—given that the framework is undefined—there are general benefits and drawbacks of government involvement in this space that merit discussion.

Perhaps the most obvious benefit is that government regulation creates a level of accountability that is absent in any system of self-regulation. Instead of merely *encouraging* platforms to engage, through use of CSR or economic goals, government regulation would *mandate* that platforms curate content in a specific manner or suffer the attend-

230. *Id.*

231. *See, e.g.*, Abby K. Wood & Ann M. Ravel, *Fool Me Once: Regulating "Fake News" and Other Online Advertising*, 91 S. CAL. L. REV. 1223, 1244 (2018) (discussing how negative market externalities justify regulation in that "the market activities are platforms chasing profits without exercising gatekeeping or transparency responsibilities, and the externalities are costs borne by social media users in their roles as voters and participants in civic life").

232. *See id.*

233. *See* Klonick, *supra* note 9, at 1626; *see* Newton, *supra* note 18.

234. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020).

235. *See supra* Part II.B (discussing how numerous legislators and commentators suggested amending section 230 to limit platform immunity).

236. *See* Klonick, *supra* note 9, at 1606–07 (describing some of the constitutional challenges identified by lawmakers and courts).

ant consequence.[237] In addition, this approach would centralize governance in a comprehensive manner—all social platforms would be required to curate content to eliminate or reduce particular harms. Government regulation would vindicate a public interest for those who assert that content on social media has led to violence and harm,[238] and thus government involvement is warranted.[239] Particularly among those who argue that social platforms demonstrate a bias against conservative views, an additional argument in favor of government regulation is that social platforms should not be in complete control over the content users see.

Yet, benefits that may be achieved in a model of government regulation are overwhelmed by significant drawbacks. First and foremost is that government efforts to regulate speech online on the basis of its content would likely run afoul of the First Amendment.[240] As previously discussed, the First Amendment protects a platform's decisions in moderating the speech it allows to be posted, promotes, or declines to display.[241] A regulatory framework requiring social platforms to treat content in a particular manner would be presumptively unconstitutional, and could be justified only if it met the highest level of judicial review: strict scrutiny.[242] Of course, strict scrutiny "requires the Government to prove that the restriction furthers a compelling interest and is narrowly tailored to achieve that interest."[243] Could some categories of harmful speech decried by commentators and lawmakers survive a strict scrutiny analysis? Possibly, for speech that falls within an unprotected category of speech, such as "incitement, . . . defamation, speech integral to criminal conduct, . . . [or] true threats."[244] But there are no assurances that such a law *would* survive, given that other

237. *Id.* at 1626 (discussing how CSR and economic goals are the two main motivators for company change); *see* Brown & Peters, *supra* note 4, at 530–31 (explaining how model legislation like the Digital Millennium Copyright Act and current notice-and-takedown law in Germany show what Congress could implement to control platforms).

238. Wood & Ravel, *supra* note 231 (explaining the harms that users experience when companies fail to remove information).

239. *See* Siegel & Reich, *supra* note 198 (discussing the call for change and greater regulation).

240. Brown & Peters, *supra* note 4, at 532–33.

241. *Id.*; *see also supra* Part II.

242. *See, e.g.*, Ashcroft v. ACLU, 542 U.S. 656, 673 (2004) (striking down the COPA); Reno v. ACLU, 521 U.S. 844, 874 (1997) (striking down a portion of the CDA).

243. Reed v. Town of Gilbert, Ariz., 576 U.S. 155, 171 (2015) (quoting Ariz. Free Enter. Club's Freedom Club PAC v. Bennett, 564 U.S. 721, 734 (2011)).

244. *See* Louis W. Tompros, Richard A. Crudo, Alexis Pfeiffer & Rahel Boghossian, *The Constitutionality of Criminalizing False Speech Made on Social Networking Sites in a Post-*Alvarez*, Social Media-Obsessed World*, 31 HARV. J.L. & TECH. 65, 89–90 (2017) (noting that when "the regulated speech falls within a circumscribed category, the Court most often submits the regulation to rational basis review, a highly deferential standard under which a law is almost always upheld").

efforts to punish online speech in unprotected categories have not always been successful.[245]

Beyond First Amendment concerns, there is an ancillary threat to speech that arises from legislating around permissible speech. Requirements to remove certain speech will have the practical effect of removing any speech adjacent to the prohibited category. Particularly where the lines between permissible and impermissible speech are difficult to draw—which is true for many speech harms—incentives are stacked in favor of takedown. For example, "content moderation aimed at terrorist propaganda can sweep in news reporting, political protest, documentary footage, and more."[246] One need look no further than the DMCA for an example of overenforcement in response to a notice-and-takedown system. Although the DMCA "relieves secondary parties of monitoring duties" upon prompt removal of flagged content, it also tends to "incentivize overenforcement, especially when it is difficult or costly to evaluate whether a user's conduct is actionable."[247] For this reason, the DMCA's "regime is widely criticized for allowing 'take-down' without adequate proof of the underlying infringement."[248] The reverse outcome would be true for regulatory frameworks that would punish platforms for over-moderating content to promote (or demote) a particular political viewpoint. The risk of over-moderation in violation of the law could induce social platforms to reduce those efforts altogether.

Perhaps the greatest concern with government regulation is that it could be abused and employed as a political tool aimed at advancing the administration's interest. The Trump administration's May 28, 2020 executive order serves as an example. Just days after Twitter added a fact-check label to two of former-President Trump's misleading tweets about mail-in voting, he reacted by developing a framework for regulation that, if not complied with, could ostensibly result in losing section 230 immunity. Despite the fact that the order was rife with legal hurdles, the point was to send a clear message to social platforms: any effort to limit or frame the President's speech, even when false and potentially damaging, will be met with aggressive legal action. This type of abuse is the ultimate concern with vesting regulatory authority for content on social platforms with the government. Indeed, one of the core justifications for the speech freedoms within the First Amendment was a "pervasive and deep-seated mistrust of gov-

245. *See, e.g.*, *Ashcroft*, 542 U.S. at 673 (holding that the COPA was likely unconstitutional because it categorically banned the distribution of content harmful to minors to anyone under seventeen for commercial purposes).

246. Llanso, *supra* note 214.

247. Madeline Byrd & Katherine J. Strandburg, *CDA 230 for a Smart Internet*, 88 FORDHAM L. REV. 405, 431 (2019); Reid, *supra* note 53, at 105 ("In this haze, ISPs are incentivized to over-block and err on the side of removing content—including lawful content.").

248. Byrd & Strandburg, *supra* note 247, at 431 (2019).

ernment."[249] Although the online marketplace of ideas contains damaging and dangerous speech, allowing the government to control that marketplace is a greater threat.[250]

Indeed, the Internet, and in particular social platforms, have been viewed as democratic frontiers: places where everyone has the opportunity to speak and have access to unfiltered ideas. It was idealized as a place where "[n]ews and information would no longer be mediated by newspaper editors, television producers and other gatekeepers. Instead, social media would allow direct access to individual voices in a feed custom-built by the user."[251] Allowing government regulation would unravel the promise of the Internet as a medium for the free exchange of thoughts and ideas.[252]

Finally, government regulation of social platforms likely comes with great administrative costs for both the social platforms and the government. Even where platforms utilize algorithms to make initial content determinations, some level of human content moderation—arguably a significant level—will likely be necessary, given the complexity and nuances of speech. Large social media companies, like Facebook and YouTube, *may* easily bear these administrative costs, but these costs will be challenging if not impossible for emerging and smaller platforms to meet. Certainly, high costs of complying with regulations would discourage new platforms from entering the market, in turn stifling competition and encouraging monopolies.

Given the barriers to government regulation and the insufficiency of self-regulation, it is necessary to find another solution to address the significant concerns related to the proliferation of harmful speech online.

3. Industry Governance: Content Moderation That Is Just and Right

A regulatory option at the industry level presents an opportunity to avoid many of the challenges inherent in the decentralized and government-led frameworks. Among the industries that use such a model, self-regulatory councils ("SRC")[253] generally exist either in place of

249. Ronald J. Krotoszynski, Jr., *Free Speech Paternalism and Free Speech Exceptionalism: Pervasive Distrust of Government and the Contemporary First Amendment*, 76 OHIO ST. L.J. 659, 679 (2015).

250. *Id.* ("[T]he greater threat comes not from private actions that distort the marketplace of ideas, but rather from state interventions to shape it.").

251. Peter Suderman, Opinion, *The Slippery Slope of Regulating Social Media*, N.Y. TIMES (Sept. 11, 2018), https://www.nytimes.com/2018/09/11/opinion/the-slippery-slope-of-regulating-social-media.html [https://perma.cc/5VSY-CML6].

252. *Id.*

253. In the context of this Article, "self-regulatory council" does not refer to "self-regulatory organizations," the narrowly defined term from securities laws, but is used more broadly to refer to a self-regulatory body composed of designates from companies, government, academics, and interest groups across a particular industry.

government regulation or as a form of co-governance. As an independent body with members, these organizations generally do not require governmental authority to enforce their regulations, as they have built-in enforcement mechanisms. These SRCs are often established to develop standards and protocols that promote order and efficiency across the industry and can be particularly useful in industries—like social media—where public trust is low.[254] Perhaps more importantly, when executed effectively, they can stave off impending government regulation. The Advertising Self-Regulatory Council ("ASRC") serves as a useful example of both purposes.[255]

In the late 1960s and early 1970s, public opinion about advertising shifted from positive or mixed attitudes to negative perceptions accompanied by significant mistrust of the industry. [256] At the same time, there was an increase in legislation designed to protect consumers and the executive branch, which made greater regulation of the advertising industry forthcoming.[257] To respond to the public distrust and impending new regulations—which advertisers were anxious to avoid—the industry adopted the ASRC as a self-regulation mechanism that included meaningful reforms.[258] The upshot was that there was an "inverse correlation between the rise of [the ASRC] and the diminution of criticism and government interest in advertising."[259]

The ASRC is not the only SRC that serves as a buffer between industry and government on the one side and the public on the other. The Financial Industry Regulatory Authority ("FINRA") is a nongovernmental organization that aims to protect investors and the markets by regulating both member firms and exchange markets.[260] The American Bar Association ("ABA") puts forth Model Rules of Professional Conduct that "serve as models for the ethics rules of most jurisdictions."[261] It also accredits law schools and is thus able to con-

254. *See, e.g.*, John D'Antona, *FINRA Reins in High-Risk Brokers*, MKTS. MEDIA (June 6, 2017), https://www.marketsmedia.com/finra-approves-proposals-control-high-risk-brokers/ [https://perma.cc/MF67-MJ9W] (noting that FINRA's regulatory efforts are necessary to improve public trust in the historically distrusted industry of stockbroking).

255. *Advertising Self-Regulatory Council*, TRUTH IN ADVERT., https://www.truthinadvertising.org/national-advertising-review-council-narc/ [https://perma.cc/QN9G-4GMB].

256. Eric J. Zanot, *The National Advertising Review Board, 1971–1976*, 59 JOURNALISM MONOGRAPHS 1, 6 (Feb. 1979), https://eric.ed.gov/?id=ed170745 [https://perma.cc/8SNR-B2PL].

257. *Id.*

258. *Id.*

259. *Id.*

260. *About FINRA*, FIN. INDUS. REGUL. AUTH., https://www.finra.org/about [https://perma.cc/6HPR-GAWJ].

261. *Model Rules of Professional Conduct: About the Model Rules*, AM. BAR ASS'N, https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/ [https://perma.cc/SP3H-JHD4].

trol parameters of what constitutes an acceptable legal education.[262] Other SRC examples include the National Association of Realtors and the American Medical Association. SRCs vary in both organizational structure and operation—some, like FINRA, operate under the authority of a government agency, while others, such as the ASRC, are independent from government but refer cases to the FTC.[263]

The ASRC provides a useful regulatory model for social platforms that could have widespread appeal for three main reasons. First, social platforms might be able to avoid cumbersome government regulation and potentially have not just input but some control over self-governance. Second, users of social platforms would benefit from an enhanced user experience because of clearly defined SRC policies and grievance processes. Even if the SRC standards for moderating content, removing harmful speech, and responding to content concerns were too restrictive or liberal for individual users' preferences, it would still be a significant improvement over the current systems and practices because the SRC would have clearly defined expectations and processes for platforms that users could rely on. Finally, the government would directly benefit from SRC management of content regulation for two reasons. First, the SRC would advance the government's interest in reducing harm without the problem of the government regulating speech—which would be constitutionally problematic. Second, SRC management would help conserve limited government resources by deferring regulation to an industry that is likely more capable of regulating its members than a government agency. An SRC built for social platforms can not only avoid problematic government involvement in speech determinations but can also provide the right incentive for social media companies to reduce known harms.

The Social Platform Regulatory Council ("SPRC") thus proposed could work in myriad ways, but I suggest four basic structural elements. First, and most importantly, the SPRC needs incentives to draw a high rate of voluntary industry participation. Second, the SPRC board needs to have the right composition. Third, members of the SPRC must be required to demonstrate accountability to a set of shared principles. Fourth, SPRC oversight must have teeth.

a. The Need for Voluntary Participation

The SPRC must offer benefits to participants that encourage voluntary participation, particularly considering that membership demands

262. *Frequently Asked Questions*, AM. BAR ASS'N, https://www.americanbar.org/groups/legal_education/resources/frequently_asked_questions/ [https://perma.cc/RRF5-QXD9].

263. *See Advertising Self-Regulatory Council*, *supra* note 255 (stating the ASRC "'may' refer cases of untruthful or deceptive advertising claims to the appropriate governmental regulatory authority").

that social platforms commit to meeting enhanced ethical, and not legal, obligations. The desire to avoid government regulation and improve public perception may be enough. The National Advertising Division of the Better Business Bureau ("NAD"), one of the self-regulatory units of the ASRC,[264] has a high rate of voluntary industry participation for just these reasons.[265] The NAD is charged with providing independent self-regulation by overseeing the truthfulness of advertising.[266] The NAD accepts complaints from competitors about deceptive advertising claims, and advertisers participate in the NAD's arbitration processes because to do otherwise risks both adverse publicity and referral to the FTC.[267] An added benefit is that "allowing competitor challenges transforms the self-regulatory process into one that takes place in a competitive and adversarial, rather than collusive, forum and encourages a high degree of participation."[268]

But benefits beyond avoiding government regulation and improving public perception could encourage social platform participation in the SPRC. Cooperation across the social media landscape could result in efficiencies for the industry, innovation, and the dissemination of useful information, which ultimately benefit both users and the social platforms themselves. Working together, it may be easier for platforms, particularly smaller and less-funded platforms, to manage some of the more challenging content issues. A working example of this is the Global Internet Forum to Counter Terrorism ("GIFCT"), which Facebook, Microsoft, Twitter, and YouTube founded in 2017 to further address terrorist abuses of their digital platforms by sharing information and collaborating to facilitate identifying and blocking terrorist content.[269]

Additionally, membership in a legitimate SPRC avoids concerns that a social media company's purported efforts to address harmful speech are simply window dressing. This was the case with Facebook's recently announced independent Oversight Board. Facebook developed its Board with twenty initial members to "rule on difficult content issues, such as whether specific Facebook or Instagram posts constitute hate speech. Some of its rulings will be binding; other will

264. *The National Advertising Review Council Is Now the Advertising Self-Regulatory Council*, BBB NAT'L PROGRAMS: NEWSROOM (Apr. 23), https://bbbprograms.org/archive/the-national-advertising-review-council-is-now-the-advertising-self-regulatory-council [https://perma.cc/BXL9-2NJB].

265. *National Advertising Division*, BBB NAT'L PROGRAMS, https://bbbprograms.org/programs/all-programs/national-advertising-division# [https://perma.cc/D3XX-LTQJ].

266. *Id.*

267. John E. Villafranco & Katherine E. Riley, *So You Want to Self-Regulate? The National Advertising Division as Standard Bearer*, 27 ANTITRUST 79, 79–80.

268. *Id.* at 80.

269. Llanso, *supra* note 214.

be considered 'guidance.'"[270] However, the Board has been roundly criticized as a "high-priced fig leaf"[271] and vested with no real power.[272] Instead of offering meaningful reform, the Board, which is funded entirely by Facebook itself, has been criticized as one that "will have no influence over anything that really matters in the world."[273] One reason for this concern is that the Board's purpose is largely to oversee whether Facebook's content enforcement decisions are consistent with the company's content policies and values.[274] Thus, the Board lacks oversight of the company's content policies and values themselves, and instead simply issues just advisory opinions on policy.[275] In addition, because Facebook hand selected the Board, "it risks becoming stacked with members who would be too deferential to the company."[276] To be effective, any oversight board would need to be made up of a diverse and well-rounded group of experts that could make decisions independent from the regulated companies.

b. The Need for a Diverse and Well-Versed Board of Experts

An objection that often arises in self-regulatory efforts is with respect to competition and anti-monopoly concerns. In an industry that is *already* perceived as having monopolistic practices, it is critically important to avoid governance that represents and advocates for the interests of large social platforms. Impartiality with respect to individual platforms is essential. To ensure meaningful efforts at self-regulation, the SPRC's leadership board must represent diverse and global interests regarding content moderation. It needs to include people who represent a variety of disciplines, including experts on free ex-

270. Margaret Sullivan, *Facebook Has a Huge Truth Problem. A High-Priced 'Oversight Board' Won't Fix It.*, WASH. POST (May 14, 2020, 11:00 AM), https://www.washingtonpost.com/lifestyle/media/facebook-has-a-huge-truth-problem-a-high-priced-oversight-board-wont-fix-it/2020/05/14/c5b53cba-95d9-11ea-9f5e-56d8239bf9ad_story.html [https://perma.cc/XR7T-RJ5B].

271. *Id.*

272. *Id.*; Dipayan Ghosh, *Facebook's Oversight Board Is Not Enough*, HARV. BUS. REV. (Oct. 16, 2019), https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough [https://perma.cc/XR3P-NXPB]; John Naughton, Opinion, *Facebook's 'Oversight Board' Is Proof That It Wants To Be Regulated – By Itself*, THE GUARDIAN (May 16, 2020, 11:01), https://www.theguardian.com/commentisfree/2020/may/16/facebooks-oversight-board-is-proof-that-it-wants-to-be-regulated-by-itself [https://perma.cc/YV9L-UXP4]; Siva Vaidhyanathan, *Facebook and the Folly of Self-Regulation*, WIRED (May 9, 2020, 2:58 PM), https://www.wired.com/story/facebook-and-the-folly-of-self-regulation/ [https://perma.cc/2ADP-WRAZ].

273. Vaidhyanathan, *supra* note 272.

274. *Oversight Board Charter*, FACEBOOK 5 (Sept. 2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/U6WS-3RV8].

275. Natasha Lomas, *Meet Facebook's Latest Fake*, TECHCRUNCH (Sept. 21, 2019, 12:56 PM), https://techcrunch.com/2019/09/21/meet-facebooks-latest-fake/ [https://perma.cc/75F6-25HG].

276. Mark Latonero, *Can Facebook's Oversight Board Win People's Trust?*, HARV. BUS. REV. (Jan. 29, 2020), https://hbr.org/2020/01/can-facebooks-oversight-board-win-peoples-trust [https://perma.cc/Z2HF-YD62].

pression, activists, lawyers, scholars, human rights leaders, linguists, business leaders, and, of course, members of social platforms. Critically, the SPRC cannot have an outsized presence of the large and powerful social media companies, such as Facebook/Instagram, YouTube/Google, or Twitter, or only the interests of those large corporations would be represented as the interests of the entire industry.

As modeled by the ASRC, the SPRC could have several subdivisions, each responsible for a discrete area where board members have particular expertise. A benefit is that the board would not need to limit itself to just decisions regarding content moderation but could potentially address concerns related to privacy, disclosure obligations, advertising, and more.

c. The Need for Accountability to a Set of Shared Principles

Without accountability to a set of shared principles fixed in a general code of conduct, an SPRC for social platforms would be meaningless. Shared principles would include commitments to reducing harmful speech online, responding to notifications of problematic content in established timeframes, responding to user appeals, and commitments to operating within certain user agreements. This would include the establishment of industry-wide baselines, allowing for some level of independence for platform decisions, but not without adequate safeguards. For example, social platforms could be required to submit plans and policies to the SPRC, which determines compliance with industry guidelines, and then provides approval. Shared principles could also differ across geographic regions and provide guidance to members beyond legal considerations, making recommendations based on cultural values and societal norms.

Accountability demands transparency so that the SPRC has a meaningful opportunity to determine whether there has been compliance. Transparency can be limited to complete disclosure with the SPRC and not the general public. Yet the shared principles and decisions reached by the SPRC must be publicly reported to maintain public confidence in the self-regulatory body.

d. The Need for Consequences to Compel Adherence

In order for the SPRC model to have a chance of success, it must have teeth. This means that instead of pure self-regulation, there likely needs to be a nexus to a government agency—which would not provide oversight, but would offer an enforcement mechanism when the SPRC deems it necessary.[277] And this enforcement mechanism

277. *See* Wood & Ravel, *supra* note 231, at 1245–46 (noting that it is not uncommon for government agencies to "provide legal backstops to the self-regulation negotiated by industry participants, along with imposition of civil or criminal penalties on violators").

must be coupled with penalties sufficient enough to induce compliance. If not from the government, outside enforcement could come from an outside organization that offers third-party oversight of the self-regulatory system. Either way, there needs to be a tangible consequence for non-compliance such that even large and well-funded social platforms would be encouraged to adhere to established standards and practices.

The SPRC should promote adjudicatory processes with procedures that encourage broad participation. One way to ensure this is to insulate participants who adhere to guidelines from enforcement actions by the government agency or third-party responsible for enforcement. The SPRC should also incorporate an appeals system for individual platform users to contribute to goals of public trust. In essence, the SPRC would need to be able to act swiftly to address concerns yet offer due process to members.

Importantly, nearly every benefit of self-regulation would apply in the SPRC model, yet the drawbacks are entirely removed. This model empowers those who know and understand the platforms' algorithmic models to work with those that research harms from speech on platforms to quickly respond and adjust to needs. It balances the need to address harms without impermissible and undesirable government involvement.

IV. Conclusion

In an industry defined by increasing technological growth, users, and user-generated content, there is no perfect solution to addressing the incidence and proliferation of speech harms. Relying on the government to regulate this space creates real threats of abuse and likely fatal First Amendment hurdles. Yet leaving control to social platforms has not worked given their track record of prioritizing economic concerns at the expense of all other considerations. At a time when public confidence in social platforms is low and government interests in pursuing regulation is high, an opportunity exists to create a self-regulatory council that can meaningfully regulate this space. But the success of an industry-wide self-governing model necessarily depends on incentives to participate, mechanisms for enforcement, and penalties for failure to comply.

Adopting this model would promote order and efficiency within the industry and appeal to citizens and lawmakers who have been calling for change while avoiding hasty and problematic government regulation. This would also have the positive effect of separating the debate about section 230 from the challenges of regulating social platforms. Such a solution avoids most of the challenges inherent in the decentralized and government-led frameworks and would likely result in efficiencies within the industry, innovation, and the dissemination of useful information, which ultimately benefit both users and the social platforms themselves.