Program on Extremism
THE GEORGE WASHINGTON UNIVERSITY

LEVERAGING CDA 230 TO COUNTER ONLINE EXTREMISM

This paper, part of the Legal Perspectives on Tech Series, was commissioned in conjunction with the Congressional Counterterrorism Caucus

ANNEMARIE BRIDY

SEPTEMBER 2019

About the Program on Extremism

The Program on Extremism at George Washington University provides analysis on issues related to violent and non-violent extremism. The Program spearheads innovative and thoughtful academic inquiry, producing empirical work that strengthens extremism research as a distinct field of study. The Program aims to develop pragmatic policy solutions that resonate with policymakers, civic leaders, and the general public.

About the Author

Annemarie Bridy is a Professor of Law at the University of Idaho. She is also an Affiliated Fellow at the Yale Law School Information Society Project and a former Visiting Associate Research Scholar at the Princeton University Center for Information Technology Policy. Professor Bridy specializes in intellectual property and information law, with specific attention to the impact of new technologies on existing legal frameworks for the protection of intellectual property and the enforcement of intellectual property rights. She has testified before Congress on the safe harbor provisions of the Digital Millennium Copyright Act and is widely published on the shifting landscape of intermediary copyright liability and online anti-piracy/anti-counterfeiting enforcement. She has also published on the copyright implications of streaming video's disruption of the television programming market and AI's incipient disruption of markets for artistic and cultural goods. She has been interviewed on IP and tech law topics for media outlets including the LA Times, Marketplace Tech Report, Bloomberg, Ars Technica, and the ABA Journal.

Professor Bridy holds a BA, *summa cum laude* and with distinction, from Boston University; an MA and a PhD from the University of California, Irvine; and a JD, *magna cum laude*, from the Temple University James E. Beasley School of Law. At the University of California, she was a Humanities Predoctoral Fellow and an Andrew W. Mellon Research Fellow in the Humanities.

The views expressed in this paper are solely those of the author, and not necessarily those of the Program on Extremism or the George Washington University.

Introduction

Current events make it plain that social media platforms have become vectors for the global spread of extremism, including the most virulent forms of racial and religious hatred. In October 2018, a white supremacist murdered 11 people at a synagogue in Pittsburgh, Pennsylvania. The shooter was an active user of the far-right social network Gab, on which he had earlier complained that a refugee-aid organization linked to the synagogue was importing foreign "invaders" to fight a "war against #WhitePeople."[1] Journalists searching the shooter's social media accounts for a motive discovered a trail of anti-Semitic posts, including notorious Jewish conspiracy memes widely shared within the far-right's online ecosystem. In March 2019, another white supremacist massacred 51 people at two mosques in Christchurch, New Zealand. Minutes before the attack, he shared links on 8chan to his Facebook page and a rambling racist manifesto.[2] Then he live-streamed the carnage to Facebook, which didn't intervene in time to keep the footage from going viral on YouTube and elsewhere.[3]

To say that extremist content online caused the Pittsburgh and Christchurch tragedies would be a gross oversimplification. At the same time, however, we must reckon with the fact that both shooters were enmeshed in extremist online communities whose members have cultivated expertise in using social media to maximize the reach of their messages.[4] YouTube's Chief Product Officer described the Christchurch massacre as "a tragedy…designed for the purpose of going viral."[5]

As offline violence with demonstrable links to online extremism escalates, regulators have made it clear that they expect the world's largest social media platforms to more actively police harmful online speech, including that of terrorist organizations and organized hate groups.[6] In the aftermath of the Christchurch shooting, New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron urged governments and tech companies to join together in the Christchurch Call, a "commitment…to eliminate terrorist and violent extremist content online."[7] As their part of the bargain, Facebook, YouTube, Twitter, and several other tech companies agreed to "[t]ake transparent, specific measures seeking to prevent the upload of

terrorist and violent extremist content and to prevent its dissemination on social media."[8]

Among US tech companies, Facebook has been the most receptive to the idea of increased government regulation. In an unusual op-ed in *The Washington Post*, Mark Zuckerberg explicitly asked Congress to tell him what to do about hate speech and terrorist propaganda on his services.[9] Such regulation would be a significant departure from past US policy concerning online speech. Since the early days of the internet, US-based online services have benefited from a policy that gives them broad discretion to set and enforce their own guidelines defining acceptable (and unacceptable) user speech. That policy, codified in section 230 of the Communications Decency Act (CDA),[10] was adopted in large part to help foster the internet's growth as a diverse forum for civic (and civil) discourse.

In recent years, section 230 has come under fire from all directions. Some critics believe its grant of broad discretion allows social media platforms to be too permissive about their users' speech. Others believe it lets them be too restrictive. At a moment of great uncertainty for the future of section 230, this article explains the positive role it can play in platforms' efforts to take greater responsibility for regulating hate speech and extremist content. I argue that the scope of immunity in Section 230 needn't be narrowed, but the statute could be productively amended to better safeguard free speech as the world's largest social media platforms turn to automated tools to comply with new, speech-restrictive European regulations.

Recent Regulatory Developments in Europe

In the European Union, the regulatory tide has turned decisively away from giving platforms broad discretion to define what counts as acceptable speech by their users. In April 2019, the European Parliament approved the "Regulation for preventing the dissemination of terrorist content online" ("Terrorist Content Regulation") requiring platforms to remove within one hour of receiving notice "material that incites or solicits

the commission or contribution of terrorist offences, or promotes the participation in activities of a terrorist group."[11] Germany's Network Enforcement Act (NetzDG) came into effect in 2018. It requires platforms to delete or block obviously illegal content, including hate speech and incitement, within 24 hours of receiving notice.[12]

Under a privately negotiated Hate Speech Code of Conduct, to which the major tech companies agreed in 2016 under pressure from the European Commission, platforms agreed to remove covered content within 24 hours of receiving notice.[13] That agreement reaches all material "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin."[14]

Given the tight timelines specified in these regimes and the massive scale at which major social media platforms operate, compliance inevitably entails deployment of algorithmic filters built on automated content-recognition ("hashing") technology.[15] The risk of over-blocking is high for mega-platforms like Facebook and YouTube, because automated systems can readily match uploaded images to stored reference files, but they cannot evaluate the context in which text and images appear.[16] For example, a system programmed to block a specific shocking or violent image cannot discern whether that image is being used legitimately as part of a news story or for the purpose of inciting violence or inflaming viewers.[17] International war crimes investigators and prosecutors have raised alarm that Facebook's algorithms for detecting terrorism-related content are blocking photographic and videographic evidence necessary to prove official violations of human rights in conflict zones.[18]

Algorithmic filtering systems also under-block.[19] They can be fooled if uploaded files are manipulated in slight ways to prevent triggering a match. The Christchurch shooting footage is a case in point. Platforms employing automated filters to block re-uploads had trouble keeping them down because so many variants were circulating, and far-right trolls who understand the limits of hashing technology were working overtime to keep the footage visible.[20] To help compensate for the limitations of automated systems, the major platforms employ legions of human content moderators to review algorithmically flagged content and to evaluate users' claims of wrongful removal.[21]

Human rights advocates are legitimately concerned that the new rules in Europe—and platforms' embrace of automated enforcement to comply with them—will negatively impact individuals' freedom of expression and freedom to receive and impart information. They worry that large quantities of lawful speech will be blocked unintentionally as platforms err on the side of caution to avoid large fines and other liability. During negotiations in the European Parliament over the terms of the Terrorist Content Regulation, civil society groups successfully advocated for changes to the proposed law that will help protect users' expressive rights under the new regime of increased liability and automated enforcement. Such provisions were weaker in the European Commission's initial text of the regulation. Thanks in large part to the efforts of these groups, the current version of the regulation contains important user- and speech-protective provisions concerning transparency, explanation, and redress.[22]

U.S. Law: The First Amendment, Section 230, and Online Speech

In the United States, unlike in the European Union, policymakers are constrained by the First Amendment when they undertake to regulate speech online.[23] Some of the most offensive speech that appears on social media platforms—for example, hate speech and many forms of targeted harassment—is protected under well-settled First Amendment case law.[24] Congress therefore cannot require Facebook to remove it. Zuckerberg's impulse to invite Congress to tell him what to do about "lawful but awful" content is understandable; he faces angry and conflicting demands from across the political spectrum. Congress, however, cannot simply wave a regulatory wand and solve Facebook's content moderation dilemma.[25]

That is not to say, however, that Facebook and other social media platforms lack the legal tools they need to target hate speech and extremist content on their services. The contrary is true. Through a combination of their own terms of service and CDA section 230, social media platforms have wide latitude to regulate users' online speech. Moreover, they are not state actors, so the First Amendment does not constrain the

types of speech they can lawfully proscribe. In other words, they can reach and remove constitutionally-protected "lawful but awful" speech. Doing so may be politically uncomfortable, as Zuckerberg appreciates, but it can unquestionably be done without any change to existing U.S. law.

Section 230 of the CDA creates space for online service providers, including social media platforms, to tailor their speech policies to the norms and expectations of their user communities. It immunizes covered providers from legal claims arising from users' speech—both speech that providers elect to host and speech they elect to block or remove.[26] Although section 230 is predominantly speech-proliferating in its effects, Congress did not want to discourage platforms from taking the initiative to block or remove user-generated content that platform operators believed to be objectionable. Accordingly, it included in section 230 a provision known as the Good Samaritan provision.

The Good Samaritan provision was intended to overrule *Stratton Oakmont v. Prodigy*, a case holding that a service provider that removed objectionable user-generated content thereby assumed editorial responsibility, and legal liability, for any unlawful content on its platform.[27] Specifically, section 230 provides immunity from suit for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider...considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[28] Section 230 thus empowers individual service providers to define the categories of content they will (not) permit users to post and publicly share on their services.

Popular Misconceptions About Section 230

Contrary to a popular misconception, Section 230 does not condition immunity on a platform's acting as a "neutral public forum."[29] Such a rule, which some conservative lawmakers advocate, would actually go a long way to *undermine* platforms' efforts to keep hate speech and extremist content offline.[30] The "otherwise objectionable" catch-

all in section 230 makes it possible for social media platforms to adopt community guidelines that ban hate speech and extremist content without loss of immunity, even if that speech is lawful in the United States. Facebook, YouTube, and Twitter have all done so, albeit not with identical definitions of the covered content.[31]

Another common misconception about section 230 is that it gives platforms a "blanket immunity" from all legal claims, allowing them to operate in a liability-free zone.[32] In fact, section 230 affords platforms no immunity from prosecution for federal crimes,[33] claims based on federal or state communications privacy law,[34] some state crimes or civil claims involving sex trafficking or prostitution,[35] or claims relating to intellectual property.[36] In addition, platforms are fully liable for any illegal content that they themselves create or develop, even in part. Section 230 immunity is thus broad but far from absolute.

The carve-out for state crimes and civil claims involving sex trafficking is the result of a 2018 amendment—the first to limit the scope of section 230 immunity since the statute was enacted in 1996. If section 230 is narrowed further, so that websites face exposure to a widening range of civil claims arising from their users' speech, it will be harder for them to manage the systemic legal risk of operating as open forums for public discussion. This is especially true for the wide universe of sites that lack the technical and financial resources of the "Big Three." If we value diversity and decentralization at the edge of the Internet, which is what section 230 was created to foster, it's important to retain section 230's immunity as a meaningful legal protection for smaller platforms.

An Effective Regulatory Tool—With Room for Improvement

Not only does the CDA permit platforms to remove material that the US government could not constitutionally prohibit, it frees them to develop and experiment with new tools for doing so, including automated technical measures. As described above, both Facebook and YouTube have implemented algorithmic filtering systems to flag and

block posts that violate their community guidelines. Although imperfect, this technology went a long way in the wake of the Christchurch shooting to help mitigate the spread of the footage. Without the automation that section 230 permits, things would have been much worse.

Platforms operating at scale and employing technical measures to moderate content face several free-speech-related challenges when confronting hate speech and extremist content. To date, they have not adequately met these challenges. One challenge is defining categories of prohibited content narrowly and precisely enough to avoid suppressing counter-speech and speech that is offensive but not plausibly harmful. A second challenge is enforcing content restrictions consistently and explaining removal decisions to users in a transparent way. A third challenge is allowing users prompt and meaningful redress for removal decisions they want to contest. In these areas, the largest platforms must do better, particularly if they are to overcome conservatives' claims of viewpoint-based bias. At least two empirical studies have shown such claims to be unfounded,[37] but unless content moderation decisions are clearly justified and explained, rumors and anecdata will persist.

To encourage the world's largest platforms to meet these challenges accountably, Congress could amend section 230 to provide protections analogous to those in the preliminarily approved version of the EU Terrorist Content Regulation. In its current form, section 230 requires platforms to exercise "good faith" in the removal of users' content, but the standard is undefined. I have argued elsewhere that US courts should interpret "good faith" in section 230 to require clarity and consistency in enforcement, and an avenue of appeal for users who believe their content has been wrongly removed.[38] It is unlikely, however, that courts will import so much substance into a legal standard that usually means little more than honesty in fact or lack of bad motive. If Congress were to amend section 230 to require transparency, explanation, and redress, US-based users would enjoy important speech protections that are embodied in the current text of the EU Terrorist Content Regulation. Such requirements could be limited to mega-platforms to avoid imposing burdens on smaller platforms whose community

guidelines and content moderation practices don't affect the expressive rights of millions or billions of people.

Conclusion

Amendments to section 230 that would require platforms to operate as "neutral public forums" would greatly undermine their efforts to keep hate speech and extremist content offline. Amendments to section 230 that would make platforms liable for illegal third-party content that they fail to detect and remove would fall hardest on platforms that are least able to bear the risk and cost of increased liability and uncertainty. For those who are concerned about economic concentration at the edge of the internet, it's worth noting that Facebook, Twitter, and YouTube are relatively well-positioned to bear the increased costs of decreased immunity, whereas start-ups and smaller sites are not.

Any amendment to section 230 should focus narrowly on protections for users of mega-platforms whose lawful speech is affected by algorithmic enforcement of community guidelines. Amendments that condition immunity for the world's largest platforms on transparency, explanation, and redress would protect users' freedom of expression and help address concerns that platforms are enforcing their community guidelines unfairly.

References

[1] J. Coaston, "How the Rise of Conspiracy Theory Politics Emboldens Anti-Semitism," *Vox*, Oct. 31, 2018.

[2] J. Weissmann, "What the Christchurch Killer's Manifesto Tells Us," *Slate*, March 15, 2019.

[3] For a detailed description of the process that played out behind the scenes at Facebook during the attack, see K. Klonick, "Inside the Team at Facebook That Dealt with the Christchurch Shooting," *The New Yorker*, April 25, 2019.

[4] *See* Z. Laub, "Hate Speech on Social Media: Global Comparisons," Council on Foreign Relations, April 11, 2019, https://perma.cc/RK24-ZUHE.

[5] E. Dwoskin and C. Timberg, "Inside YouTube's Struggles to Shut Down Video of the New Zealand Shooting—and the Humans Who Outsmarted Its Systems," *The Washington Post*, March 18, 2018.

[6] D. Ingram, "Foreign Governments Are Fed Up with Social Media - and Threatening Prison for Tech Employees," *NBC*, April 12, 2019.

[7] *See* The Christchurch Call, https://perma.cc/6XRC-GXU9. The United States declined to join the Call, citing the likelihood that its implementation will conflict with First Amendment principles. *See* T. Romm and D. Harwell, "White House Declines to Back Christchurch Call to Stamp Out Online Extremism Amid Free Speech Concerns," *The Washington Post*, May 15, 2019.

[8] The Christchurch Call, *supra* note 6.

[9] M. Zuckerberg, "The Internet Needs New Rules. Let's Start In These Four Areas.," *The Washington Post*, March 30, 2019.

[10] 47 U.S.C. § 230.

[11] European Parliament legislative resolution of 17 April 2019 on the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM(2018)0640 –C8-0405/2018 –2018/0331(COD)) ("Terrorist Content Regulation"), Recitals 9 and 13, https://perma.cc/2MKH-WG8V.

[12] W. Echikson and O. Knodt, "Germany's NetzDG: A Key Test for Combatting Online Hate," CEPS Research Report No. 2018/09, November 2018, https://perma.cc/R2FV-AG5P.

[13] The original parties to the Code were Facebook, Microsoft, Twitter and YouTube. In 2018, Instagram, Google+, Snapchat, and Dailymotion joined. *See* EU Code of Conduct on Countering Illegal Hate Speech Online, https://perma.cc/74LC-3CJB.

[14] *Id.*

[15] D. Kaye, "Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression," April 6, 2018, https://perma.cc/9XWD-7JQU.

[16] E. Engstrom and N. Feamster, "The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools," March 2017, https://perma.cc/UV5H-89SK.

[17] *Cf.*, Reporters Without Borders, "German 'Facebook Law' Creates Risk of Over-Blocking," July 10, 2017, https://perma.cc/5H29-32NB.

[18] *See* B. Warner, "Tech Companies Are Deleting Evidence of War Crimes," *The Atlantic*, May 8, 2019.

[19] S. Fussell, "Why the New Zealand Shooting Video Keeps Circulating," *The Atlantic*, March 21, 2019.

[20] *Id.*

[21] *See* T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018); S. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019).

[22] *See* EU Terrorist Content Regulation, *supra* note 10, Art. 8–11. The regulation will undergo further negotiation and potential revision before its final adoption. *See* Chloé Berthélémy,

"Terrorist Content Regulation: Successful 'Damage Control' by LIBE Committee," EDRi Blog, Apr. 8, 2019, https://perma.cc/MTD9-QAD3.
[23] *See, e.g.*, Reno v. ACLU, 521 U.S. 844 (1997) (striking down, on First Amendment grounds, provisions of the CDA intended to protect minors from "indecent" speech online).
[24] *See* E. Volokh, "No, There's No "Hate Speech" Exception to the First Amendment," *The Washington Post*, May 7, 2015.
[25] *See* D. Keller, "Who Do You Sue?," Hoover Institution Aegis Series Paper No. 19002, 2019, https://perma.cc/2UCH-48VS.
[26] The original and overarching purpose of the CDA, most of which was ultimately struck down on First Amendment grounds, was to protect children from harmful online speech. *See Reno*, *supra* note 19.
[27] *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 331 (4th Cir. 1997) ("Congress enacted § 230 to remove the disincentives to self-regulation created by the *Stratton Oakmont* decision. Under that court's holding, computer service providers who regulated the dissemination of offensive material on their services risked subjecting themselves to liability, because such regulation cast the service provider in the role of a publisher.").
[28] 47 U.S.C. § 230(c)(2)(A).
[29] For example, at a congressional hearing in 2018, Sen. Ted Cruz told Facebook CEO Mark Zuckerberg, who was then testifying, that "[t]he predicate for Section 230 immunity under the CDA is that you're a neutral public forum." Dell Cameron, "Section 230 Is the Foundation of the Internet, So Why Do Republicans Want to Change It?," *Gizmodo*, March 29, 2019.
[30] *See* A. Bridy, *supra* note 10.
[31] *See* Appendix (excerpting relevant portions of Facebook, YouTube, and Twitter Community Guidelines).
[32] D. Keats Citron and B. Wittes, "The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity," *Fordham Law Review*, 86:2 (2017).
[33] 47 U.S.C. 230(e)(1).
[34] 47 U.S.C. 230(e)(4).
[35] 47 U.S.C. 230(e)(5).
[36] 47 U.S.C. 230(e)(2).
[37] Natalie Martinez, "Study: Facebook Is Still Not Censoring Conservatives," Media Matters, April 9, 2019, https://perma.cc/9ZM4-5LKL; Natalie Martinez, "Study: Analysis Of Top Facebook Pages Covering American Political News," Media Matters, July 16, 2018, https://perma.cc/V5FS-5GZ4.
[38] A. Bridy, "Remediating Social Media: A Layer-Conscious Approach," *BU Journal of Science & Technology Law*, 28.2 (Summer 2018).