Florida Law Review

Volume 73 | Issue 6 Article 2

November 2021

Reasonableness as Censorship: Section 230 Reform, Content Moderation, and the First Amendment

Enrique Armijo
Elon University School of Law

REASONABLENESS AS CENSORSHIP: SECTION 230 REFORM, CONTENT MODERATION, AND THE FIRST AMENDMENT

Enrique Armijo[*]

Abstract

For the first time in the internet's history, revising Section 230 of the Communications Decency Act's immunity for social media platforms from liability for third-party content seems to many not just viable, but necessary. Most such calls for reform are built around the longstanding common law liability principles of duty and reasonableness, namely conditioning Section 230 liability on platforms acting reasonably to "prevent or address" third-party content that might be harmful or illegal. These reforms are finding common cause with several legislative and executive efforts seeking to compel platforms to adhere to "reasonable" or "politically neutral" moderation policies or else face increased liability for user speech. And calls for entirely new regulatory regimes for social media, some of which also call for new federal agencies to implement them, advocate for similar approaches.

This Article is the first comprehensive response to these efforts. Using the guidance of the common law to unpack the connections between reasonableness, imminence, and intermediary liability, this Article argues that these proposed reforms are misguided as a matter of technology and information policy and are so legally dubious that they have little chance of surviving the court challenges that would inevitably follow their adoption. It demonstrates the many problems associated with adopting a common-law-derived standard of civil liability like "reasonableness" as a regulatory baseline for prospective platform intermediary fault. "Reasonableness"-based Section 230 reforms would also lead to unintended, speech-averse results. And even if Section 230 were to be revised, serious constitutional problems would remain with respect to holding social media platforms liable, either civilly or criminally, for third-party user content.

* Associate Dean for Academic Affairs (through June 2021) and Professor, Elon University School of Law; Affiliated Fellow, Yale Law School Information Society Project; Faculty Affiliate, UNC-Chapel Hill Center for Information, Technology and Public Life. The Gray Center for the Study of the Administrative State at George Mason University and the Institute for Humane Studies' Free Speech and Open Inquiry Grant Program supported this Article. Thanks to Jane Bambauer, Bridget Barrett, Mike Godwin, Eric Goldman, James Grimmelmann, Rachel Kuo, Paul Levy, Alice Marwick, Amanda Reid, Matthew Sweeney, Olivier Sylvain, Alexander Tsesis, Eugene Volokh, Alex Worsnip, Adam White, Felix Wu, and Ben Zipursky for support and feedback. Thanks as well to Kathryn Romo and Cameron Capp for outstanding research assistance.

INTRODUCTION

For the first time in the relatively brief history of the internet, revising the Communications Decency Act (CDA)'s Section 230[1] to permit greater liability for social media platforms' carriage of illegal or otherwise harmful third-party content seems to many not just viable, but necessary. Whatever role Section 230 may have once played since its 1996 passage in "creat[ing] the internet we know today,"[2] in the words of one influential critique: "Today, huge social networks and search engines enable the rapid spread of destructive abuse."[3] Section 230 states that

1. Pub. L. No. 104-104, § 509, 110 Stat. 56, 137–39 (1996) (codified as amended at 47 U.S.C. § 230).

2. Jeff Kosseff, Opinion, *Section 230 Created the Internet as We Know It. Don't Mess with It*, L.A. TIMES (Mar. 29, 2019, 3:05 AM), https://www.latimes.com/opinion/op-ed/la-oe-kosseff-section-230-internet-20190329-story.html [https://perma.cc/JX9U-XH68]. *See generally* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019) (detailing the history of Section 230 of the Communications Decency Act).

3. Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 411 (2017) [hereinafter *The Internet Will Not Break*]; *see also* Danielle Keats Citron & Benjamin Wittes, *The Problem Isn't Just Backpage: Revising Section 230 Immunity*, 2 GEO. L. TECH. REV. 453, 471 (2018) (noting that ISPs and social networks encourage abuse); Danielle Keats Citron & Mary Anne Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, U. CHI. LEGAL F. 45, 68 (2020) ("The anonymity, amplification, and aggregation possibilities offered by the internet have allowed private actors to discriminate, harass, and threaten vulnerable groups on a massive scale."); Bobby Chesney & Danielle Citron, *Deep Fakes*: *A Looming Challenge for Privacy, Democracy, and*

"[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider"[4]; with a few exceptions set out elsewhere in the statute,[5] this language essentially grants immunity to platforms and websites for their hosting of user speech. This immunity for platforms from most republisher- and distributor-based liability has become untenable, so the argument goes, as those platforms are increasingly used to spread libel, harassment, terrorism, incitement, and revenge pornography, as well as to weaponize anonymous user speech.[6]

Most of these calls are built around the related and longstanding common law liability principles of duty and reasonableness.[7] The use of reasonableness in the Section 230 context would condition the liability of social media platforms, via either "judicial interpretation or legislati[ve]" amendment, on a requirement that the platforms "take[] reasonable steps to prevent or address unlawful uses of [their] services."[8] Reliance on reasonableness as a theoretical legal hook for possible intermediary liability in moderating third-party content is taking hold in Europe and the United Kingdom as well. The European Union's Digital Services Act, a draft EU-wide regulation undertaken as part of the EU's e-Commerce Directive, instructs "[v]ery large online platforms" to "put in place reasonable, proportionate and effective" content moderation practices to

---

National Security, 107 CALIF. L. REV. 1753, 1774 (2019) (discussing the ability and tendency of deep fakes to "threaten, intimidate, and inflict psychological harm" upon their victims); Alexander Tsesis, Essay, *Terrorist Speech on Social Media*, 70 VAND. L. REV. 651, 654 (2017) ("The internet is awash with calls for terrorism.").

4. *See* 47 U.S.C. § 230(c)(1).

5. The exceptions to immunity apply to certain federal criminal and intellectual property-based claims. *See id.* § 230(e).

6. *See Section 230 — Nurturing Innovation or Fostering Unaccountability?: Workshop Participant Written Submissions*, U.S. DEP'T OF JUST. 87 (Feb. 2020), https://www.justice.gov/file/1286206/download [https://perma.cc/896P-YCA3] (discussing issues with Section 230).

7. *See* RESTATEMENT (SECOND) OF TORTS §§ 4, 283 (AM. L. INST. 1965) (defining "duty" as the manner in which an "actor is required to conduct himself . . . at the risk" that his failure to do so will subject him to liability and defining the standard of conduct as "that of a reasonable man under like circumstances").

8. *The Internet Will Not Break*, *supra* note 3, at 404, 419; *see also* Ryan Hagemann, *A Precision Regulation Approach to Stopping Illegal Activities Online*, IBM THINKPOLICY LAB (July 10, 2019), https://www.ibm.com/blogs/policy/cda-230/ [https://perma.cc/8M3W-VQK9] (calling for Section 230 immunity to be "conditioned on companies applying a standard of 'reasonable care'"); L. Gordon Crovitz, *Common Law Will Finally Apply to the Internet*, NIEMANLAB (Dec. 2020), https://www.niemanlab.org/2020/12/common-law-will-finally-apply-to-the-internet/ [https://perma.cc/RS5J-T3HR] (calling for Section 230 immunity for platforms to be limited by "centuries-old," "common law" concepts of duty and reasonableness). As the IBM blog post shows, several competitors to social media companies are lobbying for changes to Section 230 as well. *See* David McCabe, *IBM, Marriott and Mickey Mouse Take on Tech's Favorite Law*, N.Y. TIMES (Feb. 4, 2020), https://www.nytimes.com/2020/02/04/technology/section-230-lobby.html [https://perma.cc/563V-CDPE].

reduce "risks" associated with their services, and would require platforms to adopt "notice-and-action" procedures.[9] Under the regulation, platforms can receive notice of allegedly illegal third-party content, and possibly be liable for failure to take action related to that content in a "timely, diligent and objective manner."[10] The UK's Department for Digital, Culture, Media & Sport and its Home Department's *Online Harms White Paper* proposed a regulatory framework for intermediary liability that relies heavily on a "duty of care," the content of which would be established and overseen by an independent regulator that would determine whether online platforms have acted reasonably with respect to third-party content.[11] The UK government is currently deciding whether the regulator enforcing this duty of care should have the power to block access to websites and "disrupt business activities" in the event of a platform's breach of the duty.[12] And the calls for new regulatory regimes for social media in the United States, with new federal agencies to implement them, advocate for similar approaches, with one former Democratic Chairman of the Federal Communications Commission calling for the establishment of a new "Digital Platform Agency" whose authority over social media companies' practices would be underpinned by a "restoration of common law principles of a duty of care."[13]

---

9. *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC,* arts. 14, 17(3), 27, COM (2020) 825 final (Dec. 15, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en [https://perma.cc/Q8GZ-RNY4].

10. *Id.*

11. JEREMY WRIGHT & SAJID JAVID, U.K. HOME DEP'T & DEP'T FOR DIGIT., CULTURE, MEDIA & SPORT, ONLINE HARMS WHITE PAPER 7, 41 (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf [https://perma.cc/9UE5-VWSG]. The government's response to comments on the *White Paper* ratified this approach, noting that under the regulations to be adopted with respect to "how the duty of care could be fulfilled," "[c]ompanies will be expected to take reasonable and proportionate steps to protect users [and t]his will vary according to the organisation's associated risk [and] size and the resources available to it." BARONESS MORGAN & PRITI PATEL, U.K. HOME DEP'T & DEP'T FOR DIGIT., CULTURE, MEDIA & SPORT, ONLINE HARMS WHITE PAPER – INITIAL CONSULTATION RESPONSE 9 (2020), https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response [https://perma.cc/2S5G-9WG3]; *see also* OLIVER DOWDEN & PRITI PATEL, U.K. HOME DEP'T & DEP'T FOR DIGIT., CULTURE, MEDIA & SPORT, ONLINE HARMS WHITE PAPER: FULL GOVERNMENT RESPONSE TO THE CONSULTATION 9 (2020), https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response [https://perma.cc/KH9M-DBKW] (providing a full government response to the *Online Harms White Paper*, including a statutory duty of care to be implemented via proposed legislation).

12. MORGAN & PATEL, *supra* note 11, at 27–28.

13. *See, e.g.*, Tom Wheeler et al., *New Digital Realities; New Oversight Solutions in the U.S.: The Case for a Digital Platform Agency and a New Approach to Regulatory Oversight*, HARV. KENNEDY SCH. SHORENSTEIN CTR. ON MEDIA, POLS. & PUB. POL'Y 6 (Aug. 2020),

Some legislative reform efforts focus on social media companies' perceived bias in their decisions as to which speakers or content to host or, to use the words of those whose access has been limited or revoked, to platforms "deplatforming" and "shadowbanning."[14] Senator Josh Hawley's June 2019 "Ending Support for Internet Censorship Act," for example, would require social media platforms with more than 30 million domestic or 300 million worldwide users and at least $500 million in global annual revenue to submit to a biannual "certification process" by the Federal Trade Commission that would ensure that the "company does not moderate information" provided by third parties "in a manner that is biased against a political party, political candidate, or political viewpoint."[15] Another Hawley bill, June 2020's "Limiting Section 230 Immunity to Good Samaritans Act," would revoke liability under 230(c)(2)'s "Good Samaritan" provision if a platform "intentionally selective[ly] enforce[s its] terms of service" to restrict certain viewpoints.[16] And Hawley's fellow senator, Ted Cruz, has used the term "neutral public forum," an undefined concept that appears nowhere in Section 230, to argue, falsely, that platforms who ban users who violate their terms of service are at risk of losing their statutory immunity.[17] An analogous bill introduced in the House, the "Stop the Censorship Act," would limit platforms' immunity for blocking content under Section 230

---

https://shorensteincenter.org/wp-content/uploads/2020/08/New-Digital-Realities_August-2020.pdf [https://perma.cc/43ZM-S99Y]; *see also* Karen Kornbluh & Ellen Goodman, *How to Regulate the Internet*, PROJECT SYNDICATE (July 10, 2019), https://www.project-syndicate.org/commentary/digital-platforms-disinformation-new-regulator-by-karen-kornbluh-1-and-ellen-p-goodman-2019-07 [https://perma.cc/GW5V-QPPW] (calling for a new federal "Digital Democracy Agency" that would regulate around issues of disinformation, privacy, and promoting local journalism); Philip M. Napoli, *What Would Facebook Regulation Look Like? Start with the FCC*, WIRED (Oct. 4, 2019, 9:00 AM), https://www.wired.com/story/what-would-facebook-regulation-look-like-start-with-the-fcc/ [https://perma.cc/8V55-CJZC] (calling for public interest-based obligations on social media platforms analogous to those imposed on broadcasters by the FCC).

14. "Deplatforming" refers to a platform's removal of users for violations of the platform's terms of service. *See, e.g.*, Rachel Kraus, *2018 Was the Year We (Sort of) Cleaned Up the Internet*, MASHABLE (Dec. 26, 2018), https://mashable.com/article/deplatforming-alex-jones-2018/ [https://perma.cc/V2GY-K632]. "Shadowbanning" refers to a platform's blocking or partially blocking of a user or their content in a way that is not readily apparent to the user. *See, e.g.*, Frank Fagan, *Systemic Social Media Regulation*, 16 DUKE L. & TECH. REV. 393, 429–30 (2018).

15. Ending Support for Internet Censorship Act, S. 1914, 116th Cong. §§ 2–3 (2019).

16. Limiting Section 230 Immunity to Good Samaritans Act, S. 3983, 116th Cong. § 2 (2020).

17. *See Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*, C-SPAN, at 1:46:25 (Apr. 10, 2018), https://www.c-span.org/video/?443543-1/facebook-ceo-mark-zuckerberg-testifies-data-protection&start=6378 [https://perma.cc/A73G-Q4TH]; *see also* Catherine Padhi, *Ted Cruz v. Section 230: Misrepresenting the Communications Decency Act*, LAWFARE (Apr. 20, 2018, 10:00 AM), https://www.lawfareblog.com/ted-cruz-vs-section-230-misrepresenting-communications-decency-act [https://perma.cc/9HHU-S77B].

for only content that is "unlawful."[18] Other proposed legislation in both chambers of Congress purports to prevent platforms from "disparate treatment" of "ideological viewpoints" as part of their content moderation practices, and also limiting liability to those platform decisions that are "objectively reasonable."[19]

Other legislative efforts at Section 230 reform purport to take more measured approaches. Some are targeted at how platforms address specific types of third-party content. These efforts also either refer to or seek to define what constitutes reasonable conduct by platforms with respect to that content. One representative proposal is Senator Lindsey Graham's 2019 Eliminating Abusive and Rampant Neglect of Internet Technologies (EARN IT) Act, which seeks to amend Section 230 to make clear that platforms and websites are not immune for the distribution of third-party child pornography over their platforms.[20] The initial version of the EARN IT Act created a "safe harbor" for a company that either (1) acts consistent with the "best practices" regarding the prevention of online child exploitation conduct developed pursuant to a Commission established by the statute, or (2) has "implemented reasonable measures" relating to online child exploitation; if the company does so, then the statute's revocation of 230 immunity would not apply.[21] Similarly, the bipartisan Platform Accountability and Consumer Transparency (PACT) Act, introduced by Senators Brian Schatz and John Thune, would require

18. Stop the Censorship Act, H.R. 4027, 116th Cong. § 2 (2019).

19. Protect Speech Act, H.R. 8517, 116th Cong. (2019); *see, e.g.*, Stop Shielding Culpable Platforms Act, H.R. 2000, 117th Cong. (2021); Abandoning Online Censorship Act, H.R. 874, 117th Cong. (2021); Safeguarding Against Fraud, Exploitation, Threats, Extremism, and Consumer Harms (SAFE TECH) Act, S. 299, 117th Cong. (2021); Curbing Abuse and Saving Expression in Technology (CASE-IT) Act, H.R. 285, 117th Cong. (2021); Protecting Constitutional Rights from Online Platform Censorship Act, H.R. 83, 117th Cong. (2021); A Bill to Repeal Section 230 of the Communications Act of 1934, S. 5020, 116th Cong. (2020); Stop Suppressing Speech Act of 2020, S. 4828, 116th Cong. (2020); Stopping Big Tech's Censorship Act, S. 4062, 116th Cong. (2020); Online Freedom and Viewpoint Diversity Act, S. 4534, 116th Cong. (2019); Curbing Abuse and Saving Expression in Technology Act, H.R. 8719, 116th Cong. (2019). FutureTense's Free Speech Project maintains an updated list of Section 230-related legislation. *See* Kiran Jeevangee et al., *All The Ways Congress Wants to Change Section 230*, SLATE (Mar. 23, 2021, 5:45 AM), https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html [https://perma.cc/G5U4-3DLY].

20. *See* EARN IT Act, S. 3398, 116th Cong. (2019) (discussion draft), https://assets.documentcloud.org/documents/6746282/Earn-It.pdf [https://perma.cc/KGZ2-C7JY]; EARN IT Act of 2020, S. 3398, 116th Cong. (2020) (amended version), https://www.judiciary.senate.gov/imo/media/doc/Graham's%20Amendment%20To%20S.3398%20-%20OLL20670.pdf [https://perma.cc/DJ5A-28WS]; Press Release, S. Judiciary Comm., Graham, Blumenthal, Hawley, Feinstein Introduce EARN It Act to Encourage Tech Industry to Take Online Child Exploitation Seriously (Mar. 5, 2020), https://www.judiciary.senate.gov/press/rep/releases/graham-blumenthal-hawley-feinstein-introduce-earn-it-act-to-encourage-tech-industry-to-take-online-child-sexual-exploitation-seriously [https://perma.cc/WJT5-7FZ9].

21. S. 3398 § 6.

platforms to take down illegal user content within twenty-four hours of receiving notice of its criminal or civil illegality, review complaints concerning content alleged to violate the platforms' terms of use within fourteen days, and publish their content moderation policies (including takedown-request and complaint-tracking policies).[22] It would also revoke Section 230-based platform immunity for illegal user "content or activity," including content that a state court has previously found to be defamatory, that the platform had knowledge of and did not remove within the relevant period.[23]

And these efforts are not limited to legislators or academics. In response to complaints about platforms' alleged bias against conservatives and Twitter's labeling of tweets by former President Donald Trump concerning fraudulent conduct associated with voting by mail, on May 28, 2020, the Trump administration promulgated an executive order titled "Preventing Online Censorship."[24] Among other things, the order proposed narrowing the executive branch's interpretation of Section 230 by revoking immunity for moderation decisions "not taken in good faith," asked the Federal Communications Commission to undertake a rulemaking to define what "good faith" requires, and required the Federal Trade Commission to assess whether platforms' content moderation practices are deceptive trade practices to the extent they "do not align with those entities' public representations about those practices."[25] Pursuant to the order, on June 17, the Department of Justice (DOJ) published a "Review of Section 230," which calls on Congress to "realign the scope of Section 230 with the realities

22. *See* PACT Act, S. 4066, 116th Cong. § 6 (2020). The PACT Act has since been updated and reintroduced. *See* PACT Act, S. 797, 117th Cong. (2021).

23. S. 4066 § 6.

24. Executive Order on Preventing Online Censorship, Exec. Order No. 13,925, 85 Fed. Reg. 34,079, 34,080–82 (June 2, 2020).

25. *Id.* (internal quotation marks omitted); Maggie Haberman & Kate Conger, *Trump Prepares Order to Limit Social Media Companies' Protections*, N.Y. TIMES (June 2, 2020), https://www.nytimes.com/2020/05/28/us/politics/trump-executive-order-social-media.html [https://perma.cc/G7R6-P2YH]; Brian Fung, *White House Proposal Would Have FCC and FTC Police Alleged Social Media Censorship*, CNN (Aug. 10, 2019, 8:15 AM), https://www.cnn.com/2019/08/09/tech/white-house-social-media-executive-order-fcc-ftc/index.html [https://perma.cc/79TB-G5VC]; *see also* Brian Fung, *Federal Officials Raise Concerns About White House Plan to Police Alleged Social Media Censorship*, CNN (Aug. 22, 2019, 5:27 PM), https://edition.cnn.com/2019/08/22/tech/ftc-fcc-trump-social-media/index.html [https://perma.cc/78QZ-4AJH] (reporting that FCC and FTC officials expressed concerns that such a proposal would violate the First Amendment). Though the FTC filed the rulemaking petition required by the Executive Order, after President Trump's supporters violently overtook the U.S. Capitol during Congress's electoral college certification on January 6, 2021, FCC Chairman Ajit Pai said he did not intend to move forward with a Section 230-related rulemaking prior to the end of his tenure as Chair. *See* Emily Birnbaum, *Ajit Pai Is Distancing Himself from Trump*, PROTOCOL (Jan. 7, 2021), https://www.protocol.com/ajit-pai-distancing-trump [https://perma.cc/8P6B-38SW].

of the modern internet" by, *inter alia*, revoking platform immunity if the platform continues to host third-party content that it has "actual knowledge or notice" of a court's judgment that the content "is unlawful in any respect," as well as for hosting third-party speech that "promotes terrorism."[26] Based on these recommendations, the DOJ submitted proposed legislation to Congress which would limit liability for removing or restricting third-party content to when the platform has an *"objectively reasonable belief"* that the restricted content was unlawful or falls into the following categories: "promot[es] terrorism or violent extremism;" displays self-harm; or is "obscene, lewd, lascivious, filthy, excessively violent," or harassing, which the immunity current covers.[27] In addition to limiting immunity to reasonable platform conduct, the DOJ's proposed amendment also revokes the statute's current immunity for taking down content that the platform subjectively believes is "otherwise objectionable"—replacing a subjective standard with an objectively reasonable one.[28]

In one of his last official acts in office, Trump vetoed the 2020 National Defense Authorization Act, holding up nearly $750 billion in unrelated military funding unless the Act "terminate[d]" Section 230 because the law posed a "very dangerous national security risk."[29] In response to that veto, then-Senate Majority Leader Mitch McConnell pointed to the "growing willingness on both sides of the aisle to at least reexamine the special legal protections afforded to technology companies" under Section 230 and declared that the Senate would "begin a process" to do so; to begin that process, he introduced legislation that would repeal Section 230 entirely.[30] And after Trump was banned from Twitter and several other platforms after a group of his supporters violently overtook the U.S. Capitol during Congress's electoral college

26. U.S. DEP'T OF JUST., SECTION 230 — NURTURING INNOVATION OR FOSTERING UNACCOUNTABILITY? 3–4 (2020) [hereinafter DOJ SECTION 230 REVIEW], https://www.justice.gov/file/1286331/download [https://perma.cc/FF2Z-GG2L].

27. *Ramseyer Draft Legislative Reforms to Section 230 of the Communications Decency Act*, U.S. DEP'T OF JUST. (Sept. 23, 2020) (emphasis added), https://www.justice.gov/file/1319331/download [https://perma.cc/FU49-FUB3].

28. *Id.*

29. Donald J. Trump, *Presidential Veto Message to the House of Representatives for H.R. 6395*, WHITE HOUSE (Dec. 23, 2020), https://trumpwhitehouse.archives.gov/briefings-statements/presidential-veto-message-house-representatives-h-r-6395/ [https://perma.cc/J36U-XDZT].

30. Press Release, McConnell on NDAA: "I Urge My Colleagues to Support This Legislation," (Dec. 29, 2020), https://www.mcconnell.senate.gov/public/index.cfm/pressreleases?ID=D8C20728-AB93-4D18-9A39-9A9391888F92 [https://perma.cc/M42N-RY29]; S. 5085, 116th Cong. (2020), https://www.congress.gov/116/bills/s5085/BILLS-116s5085pcs.pdf [https://perma.cc/MUN6-PCYT].

certification on January 6, 2021, he and several of his congressional supporters renewed their calls to revoke Section 230 immunity.[31]

As Senator McConnell's statement makes clear, attempts at Section 230 reform will not end with Trump's defeat in the 2020 presidential election. As demonstrated by the PACT Act and other legislation in both chambers of Congress, Section 230 reform is a bipartisan project.[32] And in the words of *Politico*, Section 230 reform is "[s]omething Trump and Biden agree on."[33] During his presidential campaign, President Joseph Biden spoke out against Section 230 immunity, calling for it to be "revoked[] immediately" because Facebook and other platforms are "propagating falsehoods they know to be false."[34] According to Biden, Facebook, its founder Mark Zuckerberg, and others "should be submitted to civil liability" for harmful speech in the same way as a conventional media company would be for republishing such speech.[35] Indeed, in many ways Biden's critiques of Section 230 are more directly aimed than Trump's. The former President's ire was motivated by Twitter's labeling of several of his tweets as misinformation, which does not implicate Section 230 at all because it is Twitter's own speech rather than third-party content.[36] As noted, however, Biden's concerns lie with platform

31. *See, e.g.*, Tony Romm & Josh Dawsey, *Trump Scrambles to Find New Social Network After Twitter Ban, as White House Prepares to Blast Big Tech*, WASH. POST (Jan. 9, 2021, 10:38 PM), https://www.washingtonpost.com/technology/2021/01/09/trump-twitter-banned-apps/#click=https://t.co/psNyf5W7Q7 [https://perma.cc/94ED-XQ7E]; Silvia Amaro, *Trump's Social Media Bans Are Raising New Questions on Tech Regulation*, CNBC (Jan. 11, 2021, 6:33 AM), https://www.cnbc.com/2021/01/11/facebook-twitter-trump-ban-raises-questions-in-uk-and-europe.html [https://perma.cc/ZU6N-DMQ9]; Patrick Phillips, *Graham Calls for Removal of Protections from Lawsuits After Twitter Bans Trump*, WCSC (Jan. 8, 2021, 9:18 PM), https://www.live5news.com/2021/01/08/graham-calls-removal-protections-lawsuits-after-twitter-bans-trump/ [https://perma.cc/2UFT-JDH4] ("Sen. Lindsey Graham said he's 'more determined than ever' to strip away" [Section 230 immunity].").

32. *See generally* Sabri Ben-Achour & Candace Manriquez Wrenn, *There's a Bipartisan Effort to Change Laws That Govern Speech on the Internet,* MARKETPLACE MORNING REP. (Sept. 28, 2020), https://www.marketplace.org/2020/09/28/internet-lability-law-section-230-social-media-twitter-facebook-congress-trump/ [https://perma.cc/XJS2-9Z3D] (explaining the differing incentives of both parties and various other players for repealing Section 230).

33. Renuka Rayasam & Myah Ward, *Something Trump and Biden Agree On*, POLITICO (Dec. 2, 2020, 8:09 PM), https://www.politico.com/newsletters/politico-nightly/2020/12/02/something-trump-and-biden-agree-on-491035 [https://perma.cc/3787-KHTH]; *see also* Marguerite Reardon, *Democrats and Republicans Agree That Section 230 Is Flawed*, CNET (June 21, 2020, 5:00 AM), https://www.cnet.com/news/democrats-and-republicans-agree-that-section-230-is-flawed/ [https://perma.cc/DQH7-UMVG] (explaining that both Trump and Biden agree that Section 230 should be dismantled, but disagree on "the why and how").

34. *Editorial Board Interview: Joe Biden*, N.Y. TIMES (Jan. 17, 2020), https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html?smid=nytcore-ios-share [https://perma.cc/9WR9-NLXU].

35. *Id.*

36. *See* Reardon, *supra* note 33.

immunity "from libel and civil suits for material posted on their sites by third parties, no matter how harmful"—the core of Section 230 immunity.[37] And it seems the Biden Administration's frustration with Facebook in particular became even more acute during the 2020 campaign due to what the campaign saw as Facebook's inaction concerning the spread of election-related, third-party misinformation and calls to violence on the platform.[38] Additionally, Biden's former Chief of Staff and current top technology advisor Bruce Reed recently wrote that Section 230's immunity "hurts our kids and is doing possibly irreparable damage to our democracy," compared the immunity to the federal statute that immunizes gun manufacturers from liability for gun crimes and violence, and called on Congress to "throw[] out Section 230 and start[] over."[39] Other former candidates for the Democratic presidential nomination have criticized Section 230 immunity as well.[40] In light of these developments, the bipartisan criticisms, and the likelihood that attempts to modify the statute will continue after the change in presidential administrations, even some of the platforms have expressed willingness to collaborate with the government on Section 230 reform.[41] For example, in testimony at a March 25, 2021 House of Representatives

37. Sue Halpern, *How Joe Biden Could Help Internet Companies Moderate Harmful Content*, NEW YORKER (Dec. 4, 2020), https://www.newyorker.com/tech/annals-of-technology/how-joe-biden-could-help-internet-companies-moderate-harmful-content [https://perma.cc/8U2L-THNA].

38. *See, e.g.*, Cecilia Kang, *Tweets from Biden Aide Show Campaign's Frustration with Facebook*, N.Y. TIMES (Nov. 11, 2020), https://www.nytimes.com/2020/11/11/technology/tweets-from-biden-aide-show-campaigns-frustration-with-facebook.html [https://perma.cc/7D5B-JXBD]; Cecilia Kang, *Facebook's Hands-Off Approach to Political Speech Gets Impeachment Test*, N.Y. TIMES (Oct. 14, 2019), https://www.nytimes.com/2019/10/08/technology/facebook-trump-biden-ad.html [https://perma.cc/QN37-9R3N].

39. Bruce Reed & James P. Steyer, *Why Section 230 Hurts Kids, and What to Do About It*, PROTOCOL (Dec. 8, 2020), https://www.protocol.com/why-section-230-hurts-kids#toggle-gdpr [https://perma.cc/CW7K-Y6LH]. President Biden's Commerce Department Secretary Gina Raimondo also told lawmakers during her confirmation hearings that Section 230 reform would be on the Department's agenda. *See* Makena Kelly, *Biden's Commerce Nominee Backs Changes to Section 230*, VERGE (Jan. 26, 2021, 1:40 PM), https://www.theverge.com/2021/1/26/22250746/biden-gina-raimondo-commerce-secretary-section-230 [https://perma.cc/FZ9H-RFLY].

40. *See Connecting the Dots: Combating Hate and Violence in America*, BETO FOR AM., https://www.courthousenews.com/wp-content/uploads/2019/08/beto-gun-plan.pdf [https://perma.cc/A6U7-XLW2].

41. David McCabe, *Tech Companies Shift Their Posture on a Legal Shield, Wary of Being Left Behind*, N.Y. TIMES (Dec. 15, 2020), https://www.nytimes.com/2020/12/15/technology/tech-section-230-congress.html [https://perma.cc/QMU8-PTZY] (stating that Facebook, Twitter, Reddit, and other platforms have said that they are open to discussing reforms in light of threats to make changes from both Democrats and Republicans). The efforts to modify Section 230 are part of a larger willingness by elected officials and regulators to take on social media platforms and question their power and reach, as also manifested by the Justice Department and several states' antitrust-related actions against Google and Facebook. *See id.*

hearing on the role of social media disinformation in the January 6 insurrection at the Capitol, Mark Zuckerberg proposed legislative reforms to Section 230 on Facebook's behalf that would require platforms "to demonstrate that they have systems in place for identifying unlawful content and removing it" before receiving immunity for third-party content.[42]

This Article argues that these regulatory efforts are misguided as a matter of technology and information policy, and so legally dubious that they have little chance of surviving the court challenges that would inevitably follow their adoption. Despite its appealing common law pedigree, reasonableness is a poor fit for Section 230 reform and would lead to unintended, speech-averse results. And even if Section 230 were to be legislatively revised, serious constitutional problems would remain with respect to holding social media platforms liable, either civilly or criminally, for third-party user content.

Part I shows the problems associated with adopting a common law-derived standard of civil liability like reasonableness as a baseline for prospective intermediary fault. It then considers the proliferating attempts to use products liability, another common law liability theory, to find platforms more broadly liable for third-party content. Part I also discusses the particular challenges that the use of artificial intelligence (AI) presents to the task of defining reasonableness. Part II imagines a post-Section 230 world and demonstrates how the First Amendment would remain a significant impediment to government efforts to regulate content moderation practices. Finally, Part III examines those narrow areas in which regulatory interventions might remediate harms caused by third-party content on social media.

I. COMMON LAW RIGHTS AS REGULATORY WRONGS

Several Section 230 reform efforts attempt to draw from the longstanding common law liability principle of reasonableness. Others seek to cast websites and platforms as products to establish liability under existing principles of products liability. Still others seek to distinguish the way some platforms moderate content—namely through the use of AI—

42. *See Joint Hearing Before the H. Subcomm. on Consumer Prot. & Commerce and the H. Subcomm. on Commc'ns & Tech. of the H.R. Comm. on Energy and Commerce*, 117th Cong. 7 (2021) [hereinafter *Joint Hearing*] (testimony of Mark Zuckerberg, CEO, Facebook, Inc.), https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Wstate-ZuckerbergM-20210325-U1.pdf [https://perma.cc/MBB8-36TQ]. Zuckerberg's approach is both self-serving and bad information law policy. As discussed in detail *infra* at Section I.A, an approach that sets the immunity level at practices the largest platforms are most able to reach will have the effect of entrenching those platforms at the expense of new entrants. Though Zuckerberg qualified his proposal by saying adequate content moderation systems should be "proportionate to platform size," "proportionality," like "reasonableness," is an ambiguous term that will invite litigation and incentivize over moderation. *Joint Hearing*, *supra*.

to set a hook for intermediary liability. None of these efforts, however, can adequately justify changes in the current liability rules for platforms and websites.

A. *The "Reasonableness" Problem*

The concepts of duty and reasonableness have a long pedigree in the Anglo-American common law of negligence. Succinctly stated, we owe a duty of care to those whom our conduct might foreseeably injure.[43] The content of that duty of care is defined by reasonableness.[44] When an act or omission causes another person harm, the harm-causing party's conduct will be measured by what a reasonable person would have done under the circumstances.[45] In a negligence claim, a potentially liable manufacturer or service provider's conduct will be assessed based on the possible harms another hypothetical actor in that industry would have foreseen, and what precautions such an actor would have taken to avoid those harms.[46] Reasonableness thus defines the level of care the defendant owed to the plaintiff, as well as the harmed plaintiff's factual theory of the defendant's breach that gives rise to liability. If the actor in question's act or failure to act fell below the standard that a factfinder determines was reasonable, then liability for the harm caused by that act or failure to act is appropriate.[47] Furthermore, prior harms inevitably define what possible future harms are or should have been foreseeable.[48]

Across a range of domains, the government regularly adopts reasonableness-based liability standards as part of its regulatory regimes. In principle, the government holding regulated entities to a duty of reasonable conduct as a condition of their operations is not controversial. For example, the Federal Trade Commission uses standards of unreasonableness in defining its "unfair and deceptive practices" authority.[49] In its promulgation of new car safety standards, the National Highway Traffic Safety Administration is statutorily required to consider

43. *See* RESTATEMENT (SECOND) OF TORTS §§ 4, 302 (AM. L. INST. 1965); DAN B. DOBBS ET AL., THE LAW OF TORTS § 127 (2d ed.), Westlaw (database updated June 2020) ("[T]he reasonable person exercises care only about the kinds of harm that are foreseeable to reasonable people . . . [and] risks that are sufficiently great to require precaution." (footnote omitted)).

44. RESTATEMENT (SECOND) OF TORTS § 285 (AM. L. INST. 1965).

45. *See id.* § 285 cmt. d.

46. *See id.* § 295A cmt. b ("If [an] actor does what others do under like circumstances, there is at least a possible inference that he is conforming to the community standard of reasonable conduct . . . .").

47. *See id.* §§ 282, 284, 285 cmt. h.

48. *See id.* § 285; DOBBS ET AL., *supra* note 43, § 159 (stating that foreseeability depends greatly on what the "defendant actually knew or . . . should have known").

49. Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1352 (2020) (citing Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 643 (2014)).

whether a proposed standard is reasonable.[50] Likewise, financial regulation's "rules [that] defin[e] the business of banking or ensur[e] that those institutions are safe and sound . . . turn[] on a variety of reasonableness inquiries," such as legal obligations around investor disclosures, public offering-related due diligence, and stock exchange investment standards.[51] Such uses of what this Article will call regulatory reasonableness "ensure that the one applying the law (be it legal actor or judge[, or regulator] is being guided in a manner that requires the exercise of judgment, not simply the identification of a clear-cut attribute" as with other legal rules.[52] Additionally, under the common law doctrine of negligence per se, if a plaintiff suffers a harm as a result of noncompliance with one of these regulatory reasonableness standards, in the absence of preemption, the plaintiff can often point to the noncompliance as evidence of breach of duty in a civil negligence suit.[53]

Most importantly for the present discussion, the common law of defamation states that "one who delivers or transmits defamatory matter published by a third person is subject to liability if, but only if, he knows or has reason to know of its defamatory character"; "reason to know" for purposes of the republication rule in effect means reasonableness.[54] But the unique dynamic of social media platforms—where the entity to be regulated moderates the expressive content of third parties, and that moderation is the conduct to be regulated under many of the aforementioned reform proposals—does not fit as well with reasonableness as a theoretical basis for liability.

Prospective liability based on unreasonable conduct in tort law incentivizes careful behavior, both in our interactions with others generally and in the manufacturing of products with which others will interact specifically.[55] Such prospective liability, whether imposed by tort law or a regulatory regime, has provided some degree of reliability in industries where all the entities produce similar products—say, for example, pharmaceuticals, motor vehicle production, and health care—

50. 49 U.S.C. § 30111(b)(3) (2012).

51. David Zaring, *Rule by Reasonableness*, 63 ADMIN. L. REV. 525, 543–46, 544–45 nn.84–92 (2011) (noting several statutes and regulatory authorities that apply a reasonableness standard to a range of conduct and enforcement actions in the financial sector).

52. Benjamin C. Zipursky, *Reasonableness in and out of Negligence Law*, 163 U. PA. L. REV. 2131, 2138–39, 2146 (2015).

53. RESTATEMENT (SECOND) OF TORTS § 288B(1) (AM. L. INST. 1965).

54. RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. L. INST. 1977); RESTATEMENT (SECOND) OF TORTS § 12(1) (AM. L. INST. 1965) (defining "reason to know" as the actor "ha[ving] information from which a person of reasonable intelligence . . . would infer that the fact in question exists, or that such person would govern his conduct upon the assumption that such fact exists").

55. *See* DOBBS ET AL., *supra* note 43, § 14. As discussed *infra* notes 71–73 and accompanying text, not all transmission or facilitation of a third party's defamatory statement constituted republication under the common law.

since reasonableness gives regulated entities a standard to identify and comply with. These industries also have high barriers to entry: a new firm cannot just start building cars or producing drugs without deep market knowledge. The level of sophistication required of new entrants in most large, multinational manufacturing industries thus makes it relatively easy, or at least straightforward, for those entrants to comply with standards of reasonableness imposed by private law or public regulation.

This is not at all true with respect to internet companies that host speech. Social media platforms like Facebook, YouTube, Twitter, 4chan, Grindr, Tinder, and Reddit all host third-party content, but so do Wikipedia, Dropbox, Amazon, Yelp, LinkedIn, Tumblr, and, in their online comments sections, the *New York Times* and the *Washington Post*. With a few statutory exceptions not discussed in this Article, all of these companies enjoy immunity for third-party content under Section 230,[56] but if that immunity was replaced with a duty to act reasonably, liability would then depend on a court, jury, or agency's assessment of the reasonableness of their conduct with respect to the third-party content.[57] These companies are neither comparable in the kinds of third-party content that they host nor in their capacity to moderate that content. Determining the reasonableness baseline for a particular practice is incredibly difficult when there is such a range of different approaches within that practice. Given that challenge, courts and juries will default to the stated operating procedures and content moderation practices of existing social media companies—namely those of the largest ones—to define what is reasonable or not.

Larger platforms are better able, as a matter of technological sophistication and available capital, to adopt more holistic and responsive content moderation regimes, including those that use AI (as discussed in more detail below). Smaller or start-up platforms will lack the resources to adopt such standards, preventing their development relative to incumbents.[58] The result of adopting a reasonableness standard will thus likely be the very state of affairs that many of those advocating for the change most want to avoid—an entrenching of the largest social media companies as hosts of third-party speech, an increase in their power over

56. *See* 47 U.S.C. § 230(c)(1) (2018).

57. *See* RESTATEMENT (SECOND) OF TORTS § 285 cmt. d (AM. L. INST. 1965) (stating that the reasonableness standard applies "[i]f the standard of conduct is not fixed by reference to a legislative enactment").

58. Eric Goldman, *Want to Kill Facebook and Google? Preserving Section 230 Is Your Best Hope*, BALKINIZATION (June 3, 2019, 9:30 AM), https://balkin.blogspot.com/2019/06/want-to-kill-facebook-and-google.html [https://perma.cc/4GG9-TB6Q] (arguing that because of Section 230, "startups do not need to replicate Google's or Facebook's extensive and expensive content moderation operations, nor do they need to raise additional pre-launch capital to defend themselves from business-crippling lawsuits over third-party content").

what we see and read, and a choking off of the potential alternatives to those platforms before they can even begin to compete.

A comparison to an analogous industry will demonstrate the problem. Uber has begun using geolocation tracking of its drivers to better ensure the safety of its passengers.[59] It is easy to see how such a technology might be helpful for the company to intervene if passengers are placed in danger by drivers. When a passenger is injured and claims that the harm was caused by Uber's failure to use geolocation to avoid said harm, the availability of the technology will be relevant to the decision as to whether Uber acted reasonably in supervising the driver (in addition to the more conventional evidence of direct negligence concerning hiring and supervision claims, such as the employer's efforts as to background checks, criminal records checks, drug testing and the like). After a negligence claim is brought and Uber is found to have owed a duty to the passenger bringing it, the question then becomes whether Uber's conduct sets the floor for what constitutes reasonable conduct by all ride-sharing services more generally when it comes to avoiding foreseeable harms caused to their passengers by their drivers. It is not at all difficult, in other words, to imagine a jury finding it unreasonable for *any* ride-sharing service to fail to use a risk-avoidance technology like that developed and used by Uber. This is particularly true given how susceptible juries are to hindsight-related biases and heuristics, which cause them to find an accident that has already occurred to have been more foreseeable at the time the liable party should have taken actions to prevent it.[60] Accidents and assaults that occur during Uber rides make similar risks of harm more foreseeable to other ride-sharing services, including new entrants, and thus create a duty to design and use technology to minimize those risks. If the reasonableness baseline does in fact develop in this way, the effect will be to entrench Uber as against newer ride-share startups. The same entrenchment will occur if innovation around content moderation is used to determine reasonable conduct. The result of all this, even when juries are instructed against using hindsight bias when assessing reasonableness

59. *See, e.g.*, Joseph Carino, *Uber Driver Tracking and Telematics*, GEOTAB (Jan. 15, 2018), https://www.geotab.com/blog/uber-driver-tracking/ [https://perma.cc/6ZKW-P28G]; Mary Wisniewski, *Uber Says Monitoring Drivers Improves Safety, but Drivers Have Mixed Views*, CHI. TRIB. (Dec. 19, 2016, 7:00 AM), https://www.chicagotribune.com/news/breaking/ct-uber-telematics-getting-around-20161218-column.html [https://perma.cc/Q6XM-YVGT] (discussing Uber's use of telematics technology to track driver safety).

60. *See, e.g.*, John E. Montgomery, *Cognitive Biases and Heuristics in Tort Litigation: A Proposal to Limit Their Effects Without Changing the World*, 85 NEB. L. REV. 15, 17–25 (2006) ("The[] knowledge that an event has occurred or that a bad result has been reached biases [juries] toward finding that the event or result was more foreseeable than if viewed objectively and without prior knowledge of the bad result. . . . [K]nowledge of an outcome makes it difficult for an observer to set aside that knowledge when asked to assess the factors which affect the outcome.").

and foreseeability, would be de facto strict liability for platforms' facilitation of harmful third-party speech.[61]

Moreover, the torts system from which the reasonableness standard comes is not as well-equipped to address potential intermediary liability, where a third party's conduct is the primary cause of the complained-of harm. Deciding how to design and manufacture a car or a drug is within the manufacturer's control. This manufacturer can design around the foreseeable harms associated with a particular design, manufacturing process, or warning to the extent possible. To be sure, multiple parties can be liable for a single harm in some negligence cases—the concerted action doctrine permits aiding-and-abetting-like liability when the primary tortfeasor's harm-causing conduct is "substantial[ly] assiste[d]" by another party's conduct or pursuant to a common plan.[62] And sometimes a product manufacturer can be held partially liable for harms caused by foreseeable misuses of their products by third parties.[63] But the general common law rule with respect to reasonableness is that individuals are liable for harms that their unreasonable conduct *directly* causes to other parties to whom they owe a *direct* duty of care.[64]

To take one superficially similar example of multiple parties' conduct causing a harm, if a premises owner is sued for a harm caused by a third party on the premises, the theory of liability is that the owner acted unreasonably as to the third party, violating the duty that the owner owes to all parties that are foreseeably on the owner's premises.[65] Websites and social media platforms operate very differently. To say that a social media platform owes a duty to act reasonably with respect to its users is to say it owes a duty to anyone who may be spoken of on the platform by third parties—that is, not just its users, which in the case of Facebook literally numbers in the billions[66]—but the entire world. The duty to act reasonably does not extend that far.

---

61. *Id.* at 28.

62. RESTATEMENT (SECOND) OF TORTS § 876 (AM. L. INST. 1965).

63. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 17 cmts. c–d (AM. L. INST. 1998).

64. *See* RESTATEMENT (SECOND) OF TORTS §§ 4, 302 cmt. a (AM. L. INST. 1965).

> [L]iab[ility] depends upon whether [the defendant's] breach of duty results in an injury to someone to whom the duty is owing in such a manner as to make the breach of the duty a legal cause of the injury, and this depends upon the course of events subsequent to the actor's breach of his duty, a matter over which the actor has no effective control . . . .

Id. § 4 cmt. a.

65. *See id.* § 344 cmts. d, f.

66. J. Clement, *Social Media & User-Generated Content*, STATISTA (Sept. 10, 2021), https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [https://perma.cc/C7E6-PM2Y].

Other aspects of the common law of reasonableness make it a poor fit for expanding intermediary liability for social media platforms. For liability purposes, negligence law has long distinguished misfeasance, the unreasonable failure to reduce or eliminate a foreseeable risk of harm created by one's own conduct, from nonfeasance, a failure to warn or protect someone from a risk of harm caused by someone else, which—in the absence of some other duty-creating doctrine—cannot constitute unreasonable conduct.[67] In cases involving multiple actors that are arguably involved in causing a single harm, common law doctrines like duty, proximate cause, and superseding acts are intended to place liability on the actor that is more blameworthy with respect to that harm, and cut off liability for the actor that is less blameworthy.[68] In other words, these common law doctrines place a limit on liability for mere nonfeasance with respect to the misfeasance of the true harm-causing party.[69]

A claim that a host or other platform has failed to take down the allegedly harmful speech of another party sounds more in nonfeasance than misfeasance.[70] The misfeasance/nonfeasance distinction in the speech context was recognized by the common law of republication liability's distinction between "publishers" on the one hand and "distributors" on the other. "Publishers," such as newspapers and magazines, can be liable for republishing defamatory statements because their review of such statements prior to republication means their republications are made with knowledge of the original statement's

67. These "other doctrines" include certain duty-creating relationships between the non-acting and harmed party, or whether the non-acting party's failure to act is a discontinuance of her own rescue of the harmed party. *See* DOBBS ET AL., *supra* note 43, §§ 314–30.

68. Paul F. Macri, *How the Law Court Uses Duty to Limit the Scope of Negligence Liability*, 53 ME. L. REV. 503, 504, 507 (2001).

69. *Id.*

70. Benjamin C. Zipursky, *The Monsanto Lecture: Online Defamation, Legal Concepts, and the Good Samaritan*, 51 VAL. U. L. REV. 1, 19–21 (2016). This argument assumes that the nonfeasance/misfeasance distinction applies to online defamation at all, and Zipursky also argues that the broad judicial interpretations of Section 230's immunity have blocked the common law from meaningfully reaching that question. *Id.* at 40; *see also* U.S. DEP'T OF JUST., *Section 230 Workshop – Nurturing Innovation or Fostering Unaccountability?*, YOUTUBE (May 18, 2021), https://www.youtube.com/watch?v=Jmz7xwcUPdo [https://perma.cc/B4VS-B4H6]. The statutory interpretation-based question of why Congress would have used a *conditional* term like "Good Samaritan" if it intended to confer a mostly *unconditional* immunity for platforms' content moderation decisions remains debated in internet law scholarship and case law. *See, e.g.*, Chi. Laws. Comm. for C.R. v. Craigslist, Inc., 461 F. Supp. 2d 681, 697 (N.D. Ill. 2006) ("[I]t seems rather unlikely that, in enacting the CDA and in trying to protect Good Samaritans from filtering offensive conduct, Congress would have intended a broad grant of immunity for ICSs that do *not* screen any third-party content whatsoever.").

defamatory content.[71] "Distributors," such as newsstands, libraries, or bookstores, on the other hand, could be subject to defamation liability for republication liability only if they actually knew of the original statement's defamatory character.[72] And even further removed than distributors from common law republication liability were those who "make[] available to another equipment or facilities that [the user] may use himself for general communication purposes," such as telephone companies or typewriter and loudspeaker suppliers.[73] In other words, the common law did not treat those who delivered or transmitted a defamatory statement without prior knowledge of its libelous content as publishers of said content. And those whose facilities or equipment were used to publish another's libel could not be held liable as republishers of defamation at all, absent some affirmative conduct—i.e., misfeasance—that justified finding that the equipment owner had adopted the defamatory statement as its own.

The nearly fifty-year-old *Scott v. Hull*[74] decision is one of the only cases where a plaintiff's factual theory of direct liability was the defendant's failure to take down the defamatory statement of a third party.[75] There, the court found that even though the plaintiff had given the defendant landowner notice of the defamatory statement that was graffitied on their wall and visible to the general public, the building owner could not be held liable as a common law republisher because its failure to take the statement down was mere nonfeasance.[76] The court held that failing to remove "the graffiti merely . . . after its existence was called to their attention" was not enough of a "positive act" to meet the publication requirement for common law defamation.[77] To characterize a

71. *See, e.g.*, Robert Hamilton, *Chapter 2: Defamation*, *in* KENT STUCKEY, INTERNET AND ONLINE LAW § 2.03[a] (on file with author); *see also* Loftus E. Becker, Jr., *The Liability of Computer Bulletin Board Operators for Defamation Posted by Others*, 22 CONN. L. REV. 203, 218 (1989) (stating that the principle that "there is no publication without knowledge . . . runs almost without exception through the whole of defamation law").

72. Becker, *supra* note 71, at 217–18 (citing RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. L. INST. 1979)).

73. RESTATEMENT (SECOND) OF TORTS § 581 cmt b. (AM. L. INST. 1977)); *see also* Anderson v. N.Y. Tel. Co., 35 N.Y. 2d 647, 649 (1974) (Gabrielli, J., concurring) ("It could not be said, for example, that [IBM], even if it had notice, would be liable were one of its leased typewriters used to publish a libel.").

74. 259 N.E.2d 160 (Ohio Ct. App. 1970).

75. *Id.* at 160.

76. *Id.* at 162.

77. *Id.* The *Hull* court distinguished *Hellar v. Bianco*, 244 P.2d 757 (Cal. Dist. Ct. App. 1952), an earlier case from California that applied the common law republication rule to tavern owners who failed to remove a defamatory statement about the plaintiff from their bathroom wall. *Id.* at 758–59. In *Heller*, the tavern owners' affirmative act of continuing to keep open the tavern to invitees who could see the statement that the owners refused to remove was a positive act that

social media platform's failure to take down third-party content as unreasonable thus contravenes the misfeasance/nonfeasance distinction at the heart of negligence law.

Put differently, prospective intermediary liability for third-party content will always be a facilitative cause of the complained-of harm, not a direct cause. But even under the comparative faulty system, in cases where more than one party's conduct combines to cause a harm and the system permits allocation of fault by the degree to which each party is responsible, each of those causes are direct, not facilitative.[78] Defamation actions, for example, generally do not allocate fault as between the speaker and republisher of the defamatory statement at issue.[79] The republisher's affirmative decision to disseminate the defamatory statement—again, under the common law an act of misfeasance, not of nonfeasance[80]—is a direct cause of the harm to the injured party's reputation. A social media platform, however, has not engaged in a similar affirmative act with respect to third-party content that it fails to take down.[81] The third party's defamatory or other harmful statement's *reach* may be more significant due to the platform's failure to act, but that issue goes to the secondary question of reputational damages caused by the speech's dissemination. Questions concerning the intermediary's act or failure to act do not resolve the predicate question of *liability* for the harm in question, since the publication element of defamation is met by the statement's utterance to just one person other than the plaintiff.[82]

Additionally, holding online third-party content moderation to a reasonableness standard of liability will significantly chill speech. In the absence of a mechanism by which all third-party content is screened prior to its posting (a virtual impossibility for Facebook, YouTube, or Twitter, at least), platforms will err on the side of removing any third-party speech that might be the basis for a finding of unreasonableness and thus legal liability.[83] Since the economic benefit of any single piece of user-

---

both constituted misfeasance and operated as a ratification of the defamation, such that the owners could be as directly liable as the graffitiing original defamers. *Id.* at 758. In other words, a failure to take down the statement alone was not enough for republication liability.

78. *See* RESTATEMENT (SECOND) OF TORTS § 886A cmt. h (AM. L. INST. 1979) (stating that under a comparative fault system, "percentages of fault" for all liable parties must be allocated).

79. *See* RESTATEMENT (SECOND) OF TORTS § 578 cmt. b (AM. L. INST. 1977) (stating that a republisher may be "subject to liability . . . as if he had originally published" a defamatory matter).

80. Zipursky, *supra* note 70, at 19.

81. *Cf. id.* at 21 (arguing that an ISP is more like a common carrier of another party's defamatory statement than a traditional republisher of one).

82. RESTATEMENT (SECOND) OF TORTS § 558(b) (AM. L. INST. 1977).

83. Daphne Keller, *Who Do You Sue? State and Platform Power over Online Speech* 7 (Aegis Ser. Paper No. 1902, 2019), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf [https://perma.cc/8LSE-

generated content is de minimis while potential liability as a result of that content is significant, incentives weigh heavily toward removing content that is even arguably objectionable.[84] This would result in a significantly degraded environment for speech, and, to repeat the point, a huge increase in what many current Section 230 reformers consider the greater evil—censorship of platform users' First Amendment-protected speech.

B. *The Products Liability Problem*

Generally, when the design of a manufacturer's process or product causes harm, the legal theory supporting compensation for the harmed party is one of strict liability—liability without fault.[85] Strict products liability theory, like reasonableness, is making inroads for use as a liability theory against online platforms. Some argue that (1) the platforms' designs are inherently defective with respect to how the platforms organize, post, or moderate third-party content and other information; (2) those defects cause an individual harm; and therefore (3) the platform is liable not simply vicariously, as a host for the harm-causing content, but directly, as a result of the defects in its design.[86] In the current climate, it is conceivable that states might amend their products liability statutes to permit strict liability claims against online platforms, particularly in those all-too-common instances where the actual individual or entity causing the harm is unavailable, judgment-proof, or difficult to find for purposes of a direct suit.[87] Attorneys

---

BB5C] (saying regimes that call for platforms to remove user content to avoid or minimize liability "incentivize platforms to take down speech that, while controversial or offensive, does not violate the law. Erring on the side of removing controversial speech can spare platforms legal risk and the operational expense of paying lawyers to assess content.").

84. *See, e.g.*, Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2308–14 (2014) (describing "collateral censorship" by online intermediaries); Christina Mulligan, *Technological Intermediaries and the Freedom of the Press*, 66 SMU L. REV. 157, 167, 172 (2013) (same). Indeed, the first major appellate opinion interpreting Section 230 understood this point. Zeran v. Am. Online, Inc., 129 F.3d 327, 331 (4th Cir. 1997) ("Faced with potential liability for each message republished by their services, interactive computer service providers might choose to severely restrict the number and type of messages posted.").

85. Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 4 (2018). When products liability claims are based on a dangerously defective design, however, reasonableness can play a role in assessing liability as well. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 (AM. L. INST. 1998).

86. *See, e.g.*, Ari Ezra Waldman, *The Marketplace of Fake News*, 20 U. PA. J. CONST. L. 845, 848 n.16 (2018) (arguing that because the spread of fake news is a designed-in aspect of online social network platforms, the common law of products liability for design defects could be the basis of a legal response to fake news).

87. *See, e.g.*, Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 117 (2009) (explaining that those exhibiting abusive online behavior often cover their tracks and, in any event, websites often fail to track or, after a certain period, delete, users' IP addresses); *see also*

bringing claims against platforms have increasingly embraced the theory as a work-around to Section 230.[88] Like reasonableness, however, using strict products liability as a hook for expanding potential liability for social media platforms' content moderation practices is deeply problematic.

Consistent with the current increase in skepticism toward Section 230 immunity, some courts appear to have been more receptive to products liability-related claims against online platforms for harms caused by third parties using those platforms. In 2019's *Oberdorf v. Amazon.com, Inc.*,[89] a woman who purchased a retractable dog leash from a third-party vendor on Amazon.com sued Amazon under strict products liability when she was harmed by a defect in the leash.[90] A divided panel of the Court of Appeals for the Third Circuit agreed that Amazon could be held strictly liable for her harms even though it was not the direct seller of the product, in part on the ground that the seller, a Chinese company called "the Furry Gang," could not be found.[91] The court found that the woman's failure to warn claims against Amazon were barred by Amazon's Section 230 immunity because such claims, which were rooted in the failure to "provide or to edit adequate warnings regarding the use of the dog collar," would infringe on Amazon's immunity when acting pursuant to its "publisher's editorial function."[92] But claims "premised on other actions or failures in the sales or distribution processes" such as "selling, inspecting, marketing, distributing, failing to test, or designing" those processes, would not be barred by the CDA.[93] The en banc Third Circuit vacated the panel decision, but other courts have held, consistent with the *Oberdorf* panel opinion, that Section 230 immunity may apply to posting third-party representations about products alleged to be false or misleading.[94] However, misrepresentations in the "marketing" of those products could, in theory, form a basis for strict intermediary liability under a products-based theory.[95]

David S. Ardia, *Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act*, 43 LOY. L.A. L. REV. 373, 487 (2010) (conducting an empirical study of Section 230 case law in which "41.2% of the decisions studied involved anonymous content").

88. *See, e.g.*, Jack Nicas, *Sex Trafficking via Facebook Sets Off a Lawyer's Novel Crusade*, N.Y. TIMES (Dec. 3, 2019), https://www.nytimes.com/2019/12/03/technology/facebook-lawsuit-section-230.html [https://perma.cc/B9JJ-EZTS].

89. 930 F.3d 136 (3d Cir.), *vacated and reh'g en banc granted*, 936 F.3d 182 (3d Cir. 2019) (mem.).

90. *Id.* at 140.

91. *Id.* at 147.

92. *Id.* at 153.

93. *Id.*

94. *See, e.g.*, State Farm Fire & Cas. Co. v. Amazon.com, Inc., 390 F. Supp. 3d 964, 967 (W.D. Wis. 2019).

95. *Id.*

Extending this line of reasoning, regulators and plaintiffs harmed by third-party conduct have sought to use a products liability theory to find platforms liable for the manner in which they host third-party content. Another bill of Senator Josh Hawley's, the Social Media Addiction Reduction Technology (SMART) Act, would make it unlawful for social media platforms to "automatically load[] and display[] additional content, other than music or video content that the user has prompted to play" and award badges for engaging with the social media platform that do not "substantially increase access to new or additional services, content, or functionality" on the platform.[96] Additionally, the SMART Act would affirmatively require social media companies to allow users to "set a time limit that blocks the user's own access" to the platform across all devices.[97] The SMART Act justifies such an intervention on the ground that "internet companies *design* their platforms and services to exploit brain physiology and human psychology."[98] The SMART Act and similar efforts draw from the line of products liability claims finding that the addictive level of nicotine in cigarettes constitutes a design flaw for which cigarette manufacturers might be strictly liable[99]: social media, like cigarettes, is unreasonably and dangerously addictive. The legislation, in other words, justifies regulating social media platform design based on the harms that the design exposes the users to, as products liability regulation has historically done. The theoretical hook for the SMART Act is that a social media platform is a product for purposes of strict products liability.

So far, however, most courts analyzing products liability-based claims have distinguished between platforms that place third-party products in the stream of commerce and those that host third-party speech. In *Herrick v. Grindr, LLC*,[100] Matthew Herrick, a former Grindr user, sought to hold the platform liable for false profiles of him created by a former partner that caused the user to be harassed at his home and workplace; the profiles created the false impression that Herrick was soliciting strangers for the fulfillment of sadomasochistic rape fantasies and other aggressive and violent sex.[101] Herrick argued that Grindr's app design—in particular the geolocation capability that enabled Herrick's harassers to find him at home and work based on the false profiles, the inability to detect abusive

96. SMART Act, S. 2314, 116th Cong. § 3 (2019).

97. *Id.* § 4.

98. *Id*. § 1 (emphasis added).

99. *See, e.g.*, Evans v. Lorillard Tobacco Co., 990 N.E.2d 997, 1020–21 (Mass. 2013).

100. 306 F. Supp. 3d 579 (S.D.N.Y. 2018), *aff'd*, 765 F. App'x 586 (2d Cir.), *cert. denied*, 140 S. Ct. 221 (2019).

101. Some of the fake profiles intended to create the impression that any resistance on Herrick's part would be feigned, pursuant to his interest in rape fantasies. Andrew Schwartz, *The Grindr Lawsuit That Could Change the Internet*, OUTLINE (Jan. 11, 2019, 2:02 PM), https://theoutline.com/post/6968/grindr-lawsuit-matthew-herrick [https://perma.cc/LP7G-999G].

accounts, and the failure to warn its users about abusive uses of the type he was subjected to—was a cause of his harm.[102] But the district court held that these claims were "inextricably related to Grindr's role in editing or removing offensive [third-party] content," and thus Section 230 immunity fully applied.[103] The products liability theory Herrick sought to use to get around Section 230 was unavailing; unless "the alleged duty to warn arises from something other than user-generated content," the platform could not be held liable.[104] In other words, any potential duty to warn a user concerning third-party content is precluded by Section 230. The Court of Appeals for the Second Circuit upheld the district court, and the Supreme Court declined to review the Second Circuit's decision.[105] Similarly, in the Wisconsin case of *Daniel v. Armslist*,[106] a trial court found that Armslist could potentially be liable for a mass shooter's murders because the plaintiff's claim was premised on the website's design, which made it possible for the shooter to procure the gun when he was legally prohibited from possessing one.[107] The Wisconsin Supreme Court, however, reversed, finding that the plaintiff's design-related claims all relied on the fact that Armslist was a "publisher" of third-party content for Section 230 purposes and thus immune.[108] The court held that so long as a website's design can be used for either "proper or improper purposes," Section 230 precludes the website's liability for a third-party user's subsequent use of that design for an unlawful purpose.[109]

Additionally, using strict products liability to find republication liability for social media dissemination of harmful third-party content runs afoul of a principle embedded in the First Amendment rather than Section 230: the requirement of scienter, or knowledge of one's own

102. *See* First Amended Complaint ¶¶ 108–120, at 26–27, *Herrick*, 306 F. Supp. 3d 579 (No. 1:17-CV-00932) (alleging products liability-based manufacturing and warning defect claims).

103. *Herrick*, 306 F. Supp. 3d at 588. In affirming the District Court's decision, the Second Circuit agreed with this distinction. *Herrick*, 765 F. App'x at 590 (concluding that Herrick's products liability claims were "based on information provided by another information content provider and therefore" were barred by Section 230).

104. *Herrick*, 306 F. Supp. 3d at 592.

105. Herrick v. Grinder, LLC, 765 F. App'x 586 (2d Cir. 2019), *aff'g* 306 F. Supp. 3d 579 (S.D.N.Y. 2018), *cert. denied*, 140 S. Ct. 221 (2019).

106. 913 N.W.2d 211 (Wis. Ct. App. 2018), *rev'd*, 926 N.W.2d 710 (Wis.), *cert. denied*, 140 S. Ct. 562 (2019).

107. *Id.* at 214 ¶¶ 3, 5. Armslist is a "classified advertising website similar to Craigslist" on which those seeking to buy and sell guns, including private sellers whom some states do not subject to background check requirements, can contact one another to arrange their own transactions. *Daniel*, 926 N.W.2d at 715.

108. Daniel v. Armslist, 926 N.W.2d 710, 721 (Wis. 2019) (quoting Goddard v. Google, Inc., 640 F. Supp. 2d 1193, 1197 (N.D. Cal. 2009)), *rev'g* 913 N.W.2d 211 (Wis. Ct. App. 2018), *cert. denied*, 140 S. Ct. 562 (2019).

109. *Id.* at 721.

wrongdoing, for liability for speech-related harms. In *Smith v. California*,[110] the U.S. Supreme Court held that a city ordinance that held booksellers liable for selling obscene books violated the First Amendment because it "included no element of scienter—knowledge by appellant of the contents of the book."[111] The Court found that strict liability could not be the basis for liability for carrying another's speech because "penalizing booksellers, even though they had not the slightest notice of the character of the books they sold," was incompatible with the "constitutional guarantees of the freedom of speech."[112] If booksellers could be strictly liable for obscene books, it would "impose[] a restriction upon the distribution of constitutionally protected as well as obscene literature," because "[e]very bookseller would be placed under an obligation to make himself aware of the contents of every book in his shop."[113] So too with strict liability for content-moderation practices: if third-party speech can be the basis for liability regardless of fault, platforms would err on the side of removing content that is well short of harmful or illegal because intent is by definition irrelevant when liability is strict. But as per both *Smith* and the common law of defamation republication, distributor intermediary liability cannot be strict. Products liability legal theories thus cannot support claims based on platform design decisions with respect to third-party content.

And even prior to the rise of social media, the *Restatement (Third) of Torts on Products Liability's* definition of "product" took care to distinguish between liability based on products that were "tangible personal property" that came within the law of strict products liability and the intangible "information" that can be delivered by such products.[114] As to the latter, where a

> [P]laintiff's grievance . . . is with the information, not with the tangible medium [delivering the information, m]ost courts, expressing concern that imposing strict liability for the dissemination of false and defective information would significantly impinge on free speech have, appropriately, refused to impose strict products liability in th[o]se cases.[115]

So, well before Section 230, courts and commentators found good reason to distinguish between those products for which strict liability implicated free speech values and those that did not.

110. 361 U.S. 147 (1959).

111. *Id.* at 149.

112. *Id*. at 152.

113. *Id*. at 153 (citation omitted).

114. RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 cmt. d (AM. L. INST. 1998).

115. *Id.* (discussing *Winter v. G.P. Putnam's Sons*, 938 F.2d 1033 (9th Cir. 1991), and similar cases).

As the *Herrick* case shows (at least for now), products liability theory is an unlikely end-around to Section 230, at least where courts continue to equate content moderation as publishing and editing for purposes of the statute's grant of immunity. But even if courts warmed to such approaches, they will eventually run afoul of the First Amendment's scienter-related principles as set out in *Smith*.

C. *The Algorithm Problem*

In addition to the general problems with reasonableness and strict products liability as a basis for content-moderation-derived liability described above, any new standard for intermediary liability that would replace Section 230 would have to take increasing account of the largest platforms' intent to rely more on AI in moderating content. During his congressional testimony on the Cambridge Analytica scandal, Facebook's Mark Zuckerberg referred to AI several times as the panacea for Facebook's challenges in implementing its Community Standards.[116] With four petabytes of data's worth of postings on Facebook per day,[117] human review of third-party content that potentially violates Community

116. *Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*, *supra* note 17, at 2:37:50 ("[O]ver the long term, building AI tools is going to be the scalable way to identify and root out most of th[e] harmful content" on Facebook.); *see also* Drew Harwell, *AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How*, WASH. POST (Apr. 11, 2018, 12:04 PM), https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/?utm_term=.8579661727b7 [https://perma.cc/RW7M-GZAX] (stating that Zuckerberg "referred to AI technology more than 30 times during ten hours of questioning" as the ultimate solution for platform-spoiling behavior).

In later statements, Facebook has hedged its confidence in AI's ability to solve its most difficult content moderation problems:

> While our automated tools have come a long way, they are still a blunt instrument and unable to interpret the context and intent associated with a particular piece of content. Determining a post's message is often complicated, requiring complex assessments around intent and an understanding of how certain words are being used. A person might share a news article to indicate agreement, while another might share it to condemn it. Context is critical and automated tools wouldn't know the difference, which is why relying on automated tools to identify identical or "equivalent" content may well result in the removal of perfectly legitimate and legal speech.

See, e.g., Monika Bickert, *European Court Ruling Raises Questions About Policing Speech*, FACEBOOK NEWSROOM (Oct. 14, 2019), https://newsroom.fb.com/news/2019/10/european-court-ruling-raises-questions-about-policing-speech/ [https://perma.cc/DC62-ZYUK].

117. Ankush Sinha Roy, *How Does Facebook Handle the 4+ Petabyte of Data Generated Per Day? Cambridge Analytica—Facebook Data Scandal*, MEDIUM (Sept. 15, 2020), https://medium.com/@srank2000/how-facebook-handles-the-4-petabyte-of-data-generated-per-day-ab86877956f4 [https://perma.cc/H7K2-MLW5].

Standards will never scale in a way that would satisfy Facebook's users, prospective regulators, and other constituencies.[118] Same with YouTube: It is impossible to prescreen five hundred hours of new video per minute.[119] Given this challenge, Zuckerberg discussed AI as not simply an ex post facto tool that would permit human content moderators to more quickly identify Community-Standards-infringing content, but also as a possible way to keep offending content from reaching the platform ex ante—a process that Zuckerberg argued will be faster, better, and fairer than the current ex post user/moderator notice-based system.[120] And the current international regulatory appetite for greater intermediary liability implicitly relies on the perceived feasibility of a move from ex post facto, notice-based human-moderated content moderation systems to ex ante automated ones. As Professor Hannah Bloch-Wehba observes, many of the content takedown requirements of offending third-party content being imposed on platforms by countries other than the United States will likely require platforms to filter third-party content on the upload end via the use of AI.[121]

Also, the costs to human content moderation extends to more than users who are offended, harassed, or worse. Journalistic exposés and academic studies have detailed the harms suffered by line content moderators, who are paid pittance wages to be relentlessly exposed to the worst the internet has to offer.[122] This work has caused the moderators PTSD-like trauma and to develop drug addictions, among other stress-

118. For a helpful overview of algorithmic content moderation's move from "affirmative speech control," namely promotion of content in news feeds and advertisement, into "negative speech controls," (i.e., removing, deprioritizing, or downgrading third-party content that the moderator has decided is harmful), see Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2014–16 (2019).

119. *See* H. Tankovska, *Hours of Video Uploaded to YouTube Every Minute 2007-2019*, STATISTA (Jan. 26, 2021), https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ [https://perma.cc/YAT8-TA2Y]; TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 75 (2018) ("[T]here is simply too much content and activity to conduct proactive review, in which a moderator would examine each contribution before it appeared. . . . Nearly all platforms have embraced a 'publish-then-filter' approach: user posts are immediately public, without review, and platforms can remove questionable content only after the fact.").

120. *See* Wu, *supra* note 118, at 2021 ("[S]oftware's main advantage over legal systems lies in what law would call its enforcement capacity.").

121. Hannah Bloch-Wehba, *Automation in Moderation*, 52 CORNELL INT'L L.J. 41, 69–75 (2020).

122. *See* Andrew Arsht & Daniel Etcovitch, *The Human Cost of Online Content Moderation*, JOLT DIGEST (Mar. 2, 2018), https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation [https://perma.cc/3RCQ-WEPJ].

related effects.[123] Zuckerberg apparently sees AI as a way out of this trap as well. Accordingly, by farming out the interpretation and implementation of its Community Standards to AI rather than human contract-labor reviewers, Facebook solves both its moderation problem and its moderators' problems.

As an initial matter, however, we should be skeptical of AI's ability to play a material role in content moderation, particularly with context-specific content like defamation or hate speech, with the confidence that Mark Zuckerberg communicated to Congress in 2018.[124] As a general rule, "AI may work better for images than text," as well as in areas where "there is a consensus about what constitutes a rule violation."[125] The company admitted shortly after Zuckerberg's testimony that its current AI tools only captured about thirty-eight percent of the content that it deemed hate speech in the first quarter of that year.[126] And Twitter engineers recently revealed that algorithms intended to preemptively identify and take down white supremacist-posted material would also

123. *See* David Gilbert, *Bestiality, Stabbings, and Child Porn: Why Facebook Moderators Are Suing the Company*, VICE NEWS (Dec. 3, 2019, 11:24 AM), https://www.vice.com/en_us/article/a35xk5/facebook-moderators-are-suing-for-trauma-ptsd [https://perma.cc/L49T-MAHM]; Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/UJ43-YY4W]; Jason Koebler & Joseph Cox, *The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People*, MOTHERBOARD (Aug. 23, 2018, 1:15 PM), https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works [https://perma.cc/7QRS-CX6W]; SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

124. *See, e.g.*, Neima Jahromi, *The Fight for the Future of YouTube*, NEW YORKER (July 8, 2019), https://www.newyorker.com/tech/annals-of-technology/the-fight-for-the-future-of-youtube [https://perma.cc/D8JG-4D6E] ("Machine-learning systems struggle to tell the difference between actual hate speech and content that describes or contests it."). For a summary of the deficiencies of filtering technology in the Digital Millennium Copyright Act context, see Evan Engstrom & Nick Feamster, *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools*, ENGINE (Mar. 2017), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf [https://perma.cc/S7G8-RPE4].

125. DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 63 (2019). And even image-based AI screening and filtering is much less than perfect. *See* MARY L. GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS xii, 19 (2019) ("[AI] can't always tell the difference between a thumb and a penis, let alone hate speech and sarcasm."). Some keyword-based AI filtering is also effective but not for flagging more nuanced and context-based content; "[t]his technique has not been successfully extended much past text-based profanity and slurs, (which can be based on a simple and known vocabulary)." GILLESPIE, *supra* note 119, at 98–100 (describing word filtering moderation processes).

126. Guy Rosen, *Facebook Publishes Enforcement Numbers for the First Time*, FACEBOOK NEWSROOM (May 15, 2018), https://newsroom.fb.com/news/2018/05/enforcement-numbers/ [https://perma.cc/42AY-HU9S].

sweep up tweets from Republican politicians or their supporters.[127] But putting aside technical feasibility, the important point is that AI use in content moderation complicates the use of a reasonableness standard in assessing platform intermediary liability for third-party content.

As discussed above in the context of defining reasonableness across internet companies with vastly different capacities and uses, using a regulatory-imposed duty of care to assess what constitutes reasonable platform conduct with respect to disinformation runs the risk of holding new entrants to an AI-reliant standard that no platform other than Facebook, YouTube, or Twitter could likely meet. So again, the use of a reasonableness standard could potentially have the opposite effects of what regulators intend: an entrenchment of the largest platforms, which would in turn retain and even expand the scope of harm that disinformation can cause.

Regulators hoping to regulate social media moderation practices but that are sensitive to First Amendment concerns might see a work-around in the platforms' shift to AI. From a constitutional perspective, AI-based content regulation, with its automated processes and procedures, might present a greater regulatory justification than content regulation implemented by human actions and decisions. The argument states that AI performs a function; it does not communicate.[128] Drilling further, some legal academics have argued that the move to AI-based content moderation has eroded the "distin[ction] between public functions and private functions executed by platforms," which "requires a fresh approach to restraining the power of platforms and securing fundamental freedoms" for users online.[129]

This line of thinking, however, is deeply misguided. The use of AI in content moderation does not meaningfully change the First Amendment's protections with respect to social media content moderation decisions. AI is a decision-*assistance* tool, not a decision-*making* tool.[130] The First Amendment protects human speakers and authors, not machines.[131] But

127. Joseph Cox & Jason Koebler, *Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too*, MOTHERBOARD (Apr. 25, 2019, 12:21 PM), https://www.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too [https://perma.cc/HUW4-UTJM].

128. *See* Tim Wu, *Machine Speech*, 161 U. PA. L. REV. 1495, 1521–25 (2013) (arguing that under the First Amendment functionality doctrine, AI communication tools perform tasks "unrelated to the communication of ideas" and are therefore exempt from free speech protection).

129. Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV. 857, 862–63 (2020).

130. *See* Selbst, *supra* note 49, at 1374.

131. *See* Tim Wu, *Free Speech for Computers?*, N.Y. TIMES (June 19, 2012), https://www.nytimes.com/2012/06/20/opinion/free-speech-for-computers.html [https://perma.cc/

even though the product of most algorithmic authorship is automation, all algorithms begin with human authors.[132] Even automated content moderation is simply a form of editing—"deciding [which content] to publish, withdraw, postpone or alter"[133]—a category of speech that receives full First Amendment protection.[134] Facebook's decision to remove or minimize posts that foster, to use its words, "polarization and extremism" has expressive meaning, as its moderation is a statement of its views as to the value of that category of third-party content with respect to its community of users.[135] A content-moderating algorithm, then, is just expressing the message of the individuals who wrote the code that directs the algorithm to moderate: the expressive content of the algorithm's decisions are interpretations and implementations of the platforms' First Amendment-protected terms of service.[136] The content-moderating AI that Mark Zuckerberg envisions for Facebook's future would replicate the expressive decisions of human moderators with respect to content—only faster, cheaper, and more reliably.

In addition, a deep academic literature has developed around the theme of algorithmic bias. In particular, some scholars argue that the biases and value judgments of AI's creators are embedded within AI and have deleterious effects when those algorithms are applied to members of communities that have been the object of those human-based biases and value judgments.[137] This literature necessarily relies on the presumption that algorithms are speech, since bias (even implicit bias) is

---

HE25-835J] ("To give computers the rights intended for humans is to elevate our machines above ourselves.").

132. *See, e.g.*, Stuart Minor Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1479 (2013) ("The fact that an algorithm is involved does not mean that a machine is doing the talking.").

133. Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997).

134. *See* Pittsburgh Press Co. v. Pittsburgh Comm'n on Hum. Rels., 413 U.S. 376, 391 (1973) (reaffirming the First Amendment speech "protection afforded to editorial judgment"); Mia. Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 258 (1974).

135. Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (May 5, 2021), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/32EV-3ZCZ].

136. All of the traditional theoretical justifications for the First Amendment—enabling self-autonomy, ensuring a marketplace of ideas, and facilitating democratic self-governance—also support constitutional protection for algorithmic speech. *See* Margot E. Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law*, 51 U.C. DAVIS L. REV. 589, 606 (2017).

137. *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024 (2017); Thomas Davidson et al., *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, ARXIV (2019), https://arxiv.org/pdf/1905.12516.pdf [https://perma.cc/C298-HMKZ]; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 813 (2020); Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, *in* PROC. OF THE 57TH ANN. MEETING OF THE ASS'N OF COMPUTATIONAL LOGISTICS 1668, 1668 (2019), https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf [https://perma.cc/5EK2-736X].

expressive in nature. The First Amendment protects substantive communications such as content moderation decisions and their implementation, even if those communications are expressed through artificial intelligence.

Government officials will not be able to escape First Amendment scrutiny of any efforts to regulate content moderation practices on the grounds that the moderation is automated via artificial intelligence. Modifying or eliminating Section 230's statutory immunity for republication liability, when combined with a drastic increase in AI's content moderation role, will run headlong into the argument that algorithms are First Amendment-protected speech.

II. CONSTITUTIONAL PROBLEMS IN A POST-230 IMMUNITY WORLD

Even in a world where Section 230's immunity was significantly revised or eliminated altogether, serious constitutional problems with imposing greater liability for social media platforms' hosting of harmful speech would remain.

To begin, an obvious point bears reemphasis, especially given the current regulatory appetite: *content moderation policies are protected speech.*[138] Private parties have brought dozens of cases against internet platforms complaining of the platforms' decisions to take down content.[139] In these cases, courts unanimously found that content moderation decisions are protected speech by private parties and civil liability was thus not possible.[140] There is no reason to assume that increased regulation of content moderation policies would compel a different result.

---

138. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1602, 1607–08 (2018) (stating "platforms are motivated to moderate by both of § 230's purposes: fostering Good Samaritan platforms and promoting free speech").

139. *See, e.g.*, David L. Hudson, Jr., *Free Speech or Censorship? Social Media Litigation Is a Hot Legal Battleground*, A.B.A. J. (Apr. 1, 2019, 12:05 AM), https://www.abajournal.com/magazine/article/social-clashes-digital-free-speech [https://perma.cc/3UYR-HUJF] (discussing free speech on the internet and related litigation); Adi Robertson, *Social Media Bias Lawsuits Keep Failing in Court*, VERGE (May 27, 2020, 5:43 PM), https://www.theverge.com/2020/5/27/21272066/social-media-bias-laura-loomer-larry-klayman-twitter-google-facebook-loss [https://perma.cc/BV72-C8D3] (listing numerous cases dismissed under Section 230).

140. *See* Robertson, *supra* note 139 (showing unanimous reasoning by diverse courts); Eric Goldman & Jess Miers, *Online Account Terminations/Content Removals and the Benefits of Internet Services Enforcing their House Rules*, 1 J. OF FREE SPEECH L. 191, 192 (2021) (creating dataset of judicial decisions demonstrating that "Internet services have won essentially all of the lawsuits to date brought by terminated/removed users," which shows that under current First Amendment law, "Internet services currently have unrestricted legal freedom to make termination/removal decisions").

But thanks in part to Facebook and Twitter's roles in organizing and facilitating protests around Black Lives Matter and George Floyd's murder by a police officer in Minneapolis, the interrelation between the First Amendment and social media use is coming into sharper relief. Current debates around both regulating social media—including the DOJ's efforts to amend Section 230 to hold platforms liable for failing to take down third-party speech that "promotes terrorism"[141] and law enforcement's surveillance of social media to bring incitement-based charges against social media users[142]—will require courts to directly consider the First Amendment's reach as to both platforms and their users.

A. *The Imminence Problem*

As the Supreme Court stated in 1982, the mere fact that a crime involves speech such as encouragement, solicitation, or conspiracy does not immediately trigger First Amendment review.[143] Nor is there any constitutional problem with criminal aiding-and-abetting liability where the aiding is done through the use of speech, "even if the prosecution rests on words alone."[144] But the Court has also held that the First Amendment protects most advocacy of illegal action, with one exception: advocacy that is both intended and likely to incite or produce imminent lawless action.[145] There is a significant amount of speech published on social

141. *See* DOJ SECTION 230 REVIEW, *supra* note 26, at 4.

142. *See, e.g.*, Cyrus Farivar & Olivia Solon, *FBI Trawled Facebook to Arrest Protesters for Inciting Riots, Court Records Show*, NBC NEWS (June 19, 2020, 4:26 PM), https://www.nbcnews.com/tech/social-media/federal-agents-monitored-facebook-arrest-protesters-inciting-riots-court-records-n1231531 [https://perma.cc/B3WA-MAZC] (describing the arrest and subsequent dropping of charges for incitement under the federal Anti-Riot Act based on a protestor's Facebook posts).

143. Brown v. Hartlage, 456 U.S. 45, 54–55 (1982) ("The fact that [an illegal] agreement necessarily takes the form of words does not confer upon it, or upon the underlying conduct, the constitutional immunities that the First Amendment extends to speech.").

144. United States v. Freeman, 761 F.2d 549, 552 (9th Cir. 1985) (opinion by then-Judge Kennedy).

145. District of Columbia v. Garcia, 335 A.2d 217, 223–24 (D.C. 1975). The difference between protected advocacy and unprotected solicitation has been described by one court as follows:

> [T]here is a significant distinction between advocacy and solicitation of law violation in the context of freedom of expression. Advocacy is the act of "pleading for, supporting, or recommending; active espousal" and, as an act of public expression, is not readily disassociated from the arena of ideas and causes, whether political or academic. Solicitation, on the other hand, implies no ideological motivation but rather is the act of enticing or importuning on a personal basis for personal benefit or gain.

media that directly advocates the commission of illegal activity and even violence, from incitements to riot to threats of bodily harm against individuals to calls for ethnic genocide. Many of these posts fall into the category of hate speech; some, like the manifestos posted by the perpetrators of the mass shootings in El Paso, Pittsburgh, Charleston, and New Zealand, deserve to be called much worse. But there are two significant barriers to holding such speakers liable for their speech and regulating the platforms that carry it. One is the specific intent requirement for inchoate crimes like incitement, and the other is the constitutional requirement that incitement can only be punished if the illegal acts the speaker advocates are imminent.

The primary impediment to regulating platforms' carriage of hate speech advocating violence or other criminal activity is the fifty-year-old case *Brandenburg v. Ohio*.[146] In *Brandenburg*, the Supreme Court held that, because of the First Amendment, speech advocating the use of force or legal violation could only be punished if it was intended and likely "to incit[e] or produc[e] *imminent* lawless action."[147] A common-law-derived term, both the law of assault in torts and the First Amendment doctrine of incitement have long understood imminence to essentially mean "no significant delay," or "almost at once."[148] Related areas of common law tort that also use an imminence requirement, such as the affirmative defense of necessity, where an actor seeks to have an intentional tort excused on the ground it was committed to avoid a larger harm, similarly define the term "imminence" to mean near-immediacy.[149]

The common law imposed an imminence requirement because assault as an avenue for civil liability was directly "tied to failed battery cases"—

Id. at 224 (quoting Gitlow v. New York, 268 U.S. 652, 665 (1958)); Marc Rohr, *Grand Illusion?: The* Brandenburg *Test and Speech That Encourages or Facilitates Criminal Acts*, 38 WILLAMETTE L. REV. 1, 27 (2002) (quoting *Garcia*, 335 A.2d at 224); *see also* Thomas Healy, Brandenburg *in a Time of Terror*, 84 NOTRE DAME L. REV. 655, 671 (2009) ("Criminal instruction differs from criminal advocacy in that the speaker instructs or teaches others how to commit crime instead of, or in addition to, encouraging them to do so.").

146. 395 U.S. 444 (1969) (per curiam).

147. *Id.* at 447 (emphasis added).

148. RESTATEMENT (SECOND) OF TORTS § 29(1) cmt. b (AM. L. INST. 1979). The first *Restatement of Torts* used the term "immediate," but the second *Restatement* substituted "imminent" for "immediate," in order to make it clear that the contact apprehended need not be an instantaneous one. *Id.; see also* RESTATEMENT (THIRD) TORTS: INTENTIONAL TORTS TO PERSONS § 105 cmt. e (AM. L. INST., Tentative Draft No. 1, 2015) (defining "imminent" to mean that "the contact will occur without significant delay"); DOBBS ET AL., *supra* note 43, § 39 (stating that plaintiffs must fear the battery at issue will occur "without delay unless an intervening force prevents it or the plaintiff is able to flee. Future danger, or a threatening atmosphere without reason to expect some immediate touching, in other words, is not enough." (footnotes omitted)).

149. *See* Eliers v. Coy, 582 F. Supp. 1093, 1097 (D. Minn. 1984) (finding necessity defense not available to defendant because of no "danger of imminent physical injury" justifying defendant's false imprisonment of plaintiff).

i.e., if a threatening defendant attempted to batter the plaintiff but failed to cause a harmful or offensive contact, the plaintiff could still recover if the plaintiff was aware of the defendant's attempt.[150] To put it more colloquially, a puncher with bad aim should not escape tort liability because they swung and missed but intended to punch instead of frighten. Professor Zechariah Chafee, writing in 1919, understood "the common law of incitement" to include this strict temporal connection between the threat of action and the action itself.[151] As Chafee said, the First Amendment permits punishing a speaker for "[p]olitical agitation" that "stimulate[s] men to the violation of the law . . . just before it begins to boil over" into illegal acts by listeners, and "it is unconstitutional [for government] to interfere when it is merely warm."[152]

After the aforementioned ethnic-hate-based shootings in New Zealand and El Paso, there have been several calls to hold social media platforms responsible for hosting hate speech that advocates violence or other illegal acts.[153] But there are serious problems associated with holding a republisher of incitement liable to the same degree as the initial speaker, similar to the problems identified by the courts (as discussed in Part I above) in holding the republisher of a defamatory statement equally liable as the initial defamer.[154] In *Brandenburg v. Ohio*, the government became aware of KKK leader Clarence Brandenburg's speech after a KKK rally

150. RESTATEMENT (THIRD) OF TORTS: INTENTIONAL TORTS TO PERSONS § 105 cmt. e (AM. L. INST., Tentative Draft No. 1, 2015).

151. *See* Zechariah Chafee, *Freedom of Speech in War Time*, 32 HARV. L. REV. 932, 963 (1919).

152. *Id.* at 962–63 (quoting Masses Pub. Co. v. Patten, 244 F. 535, 540 (S.D.N.Y. 1917)); *see also id*. at 967 (observing how Justice Holmes' clear and present danger test as articulated in *Schenck v. United States* "draws the boundary line very close to the test of incitement at common law and clearly makes the punishment of words for their [mere] bad tendency impossible").

153. *See, e.g.*, *Connecting the Dots: Combating Hate and Violence in America*, *supra* note 40 ("Informational service providers of all sizes, including domain name servers and social media platforms, also would be held liable where they are found to knowingly promote content that incites violence."); Makena Kelly, *Beto O'Rourke Seeks New Limits on Section 230 as Part of Gun Violence Proposal*, VERGE (Aug. 16, 2019, 1:05 PM), https://www.theverge.com/2019/8/16/20808839/beto-orourke-section-230-communications-decency-act-2020-president-democrat-background-checks/ [https://perma.cc/2QD8-4SC4].

154. *See, e.g.*, Danielle Allen & Richard Ashby Wilson, *The Rules of Incitement Should Apply to—and Be Enforced on—Social Media*, WASH. POST (Aug. 8, 2019, 4:41 PM), https://www.washingtonpost.com/opinions/2019/08/08/can-speech-social-media-incite-violence/ [https://perma.cc/5LDX-MZRX] (explaining how incitement rules are applicable to social media companies); Alexander Tsesis, *Social Media Accountability for Terrorist Propaganda*, 86 FORDHAM L. REV. 605, 619–20 (2017) ("[A] social media company that is made aware that a foreign terrorist organization has uploaded materials on its platform should be legally obligated to remove it" and "be held criminally liable to communicate the gravity of helping terrorists advance their machinations."); Richard Ashby Wilson & Jordan Kiper, *Incitement in an Era of Populism: Updating* Brandenburg *After Charlottesville*, 5 PA. J.L. & PUB. AFFS. 189, 194 (2020) (commenting on the shortfalls of incitement law).

was broadcast as part of a Cincinnati television station's report.[155] There is no indication that the prosecutors considered bringing charges against the station for airing Brandenburg's call to violence along with their charges against Brandenburg himself. To the contrary, the station's carriage of the speaker's speech was the method the government used to obtain evidence of the speech it thought to be illegal.[156]

Relatedly, analogies comparing Facebook's role in the ethnic cleansing of Rohingya in Myanmar to that of the RTLM radio station during the Rwandan genocide are fundamentally flawed.[157] In the latter case, the radio station itself was calling for and facilitating the systemic murder of the country's Tutsi population.[158] The difference, in other words, is one of intent. The publisher in the Rwanda case intended to incite violence, but that was because the publisher was also the speaker—to use a distinction from Part I, its liability was direct, not intermediary. It certainly republished speech of others' incitements as well, but the intent of the publisher and republisher in those cases was one and the same, and so coextensive liability was justified for the crimes the speech facilitated. This is not to excuse Facebook's actions and inactions around the world with respect to inciting third-party content on its platform. But the direct-versus-intermediary distinction is critical to a careful application of incitement law.

It may be so that traditional media's editing and commentary functions preclude republication liability for incitement, while social media's hosting of third-party content without modifying the contents make the possibility of intermediary liability for incitement a closer case. Traditional media often reports on past events rather than those that are about to happen; this may also be different for incitement purposes from Facebook permitting the posting of a pre-massacre manifesto. But incitement, like the other inchoate crimes, requires specific intent.[159] Unlike defamation, which can give rise to liability based on recklessness,[160] or even negligence in the case of a private person,[161] a

155. 395 U.S. 444, 445 (1969).

156. *Id.*

157. *See, e.g.*, Eric Paulsen, *Facebook Waking Up to Genocide in Myanmar*, DIPLOMAT (Sept. 21, 2018), https://thediplomat.com/2018/09/facebook-waking-up-to-genocide-in-myanmar/ [https://perma.cc/2U8G-JU2T] (comparing the media used in the Myanmar and Rwanda genocides); Timothy McLaughlin, *How Facebook's Rise Fueled Chaos and Confusion in Myanmar*, WIRED (July 6, 2018, 7:00 AM), https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/ [htpps://perma.cc/5QDZ-LVBR].

158. *See, e.g.*, Paulsen, *supra* note 157.

159. *See* Wilson & Kiper, *supra* note 154, at 206; *see also, e.g.*, State v. Dargatz, 614 P.2d 430, 437 (Kan. 1980) (finding that specific intent is required for the statutory crimes of incitement to riot and incitement to disorder).

160. RESTATEMENT (SECOND) OF TORTS §§ 580A–580B (AM. L. INST. 1979).

161. *Id.* § 580B.

social media platform cannot be liable for incitement unless it *intended*, by letting a third-party post the inciting content, to cause its users to commit imminent violent or other illegal acts.[162] In such a case, the intermediary has adopted the incitement as its own.

The imminence requirement complicates the possibility of intermediary platform liability in other ways as well. The First Amendment work that imminence does in incitement doctrine is straightforward: when a speaker riles up a mob with his words such that the mob is moved to commit bad acts immediately thereafter, even though the source of liability is speech it is nevertheless fair to find the acts and the speech analogously responsible for the harms caused by the acts, and to hold the speaker and the mob equally liable for those acts.[163] Punishing only speech likely to incite imminent unlawful activity is also justified from efficiency and deterrence perspectives. As Professor Thomas Healy writes:

> Where criminal advocacy is likely to lead to imminent lawless conduct, the government has no alternative but to criminalize the speech in the hope of deterring speakers from engaging in it[, b]ut where criminal advocacy is likely to lead to future lawless conduct [that will occur at some later point], the government can rely on police intervention, counterspeech, and the deliberation of listeners to prevent the crime from occurring.[164]

Incitement, like other inchoate crimes, is "designed to interdict a harmful chain of causation once a substantial step has been taken towards commission."[165] The primary motivation for criminalizing inciting speech is thus to prevent the crimes that the speakers would otherwise encourage from being committed in the first place.

This fundamental dynamic changes, however, when the speech the government seeks to punish or proscribe is not heard by a gathered mob but read on a screen by individuals making up a geographically diffuse audience. First Amendment doctrine can justify punishing the speaker based on the content of her speech in contravention of the general doctrinal speech-protective rule because the *context* for the speech to be punished permits a prediction that the speech will cause a listener or

---

162. *See, e.g.*, Allen & Wilson, *supra* note 154 (explaining how incitement rules are applicable to social media companies).

163. *See* Alan Chen, *Free Speech and the Confluence of National Security and Internet Exceptionalism*, 86 FORDHAM L. REV. 379, 389 (2017) ("Part of the justification for punishing the inciting speaker is that speakers in some circumstances will engage in such powerful rhetoric that it will virtually overcome the will of the listener, compelling him to engage in criminal conduct that he would not otherwise have carried out."). In other words, the decision to act illegally was the speaker's decision, not the listener's decision.

164. Healy, *supra* note 145, at 716.

165. Wilson & Kiper, *supra* note 154, at 215.

listeners to respond to its call for violence or other illegal acts.[166] "[T]he identity of the listeners and the speaker, the place, . . . the crime being advocated,"[167] and the listeners' opportunity to commit that crime in advance of any meaningful preventative police intervention—all of these factors must cut in favor of punishing the speaker in order to prevent the violent act for which the speaker advocates. As a general matter, incitement-based liability or regulation is difficult to justify when the "listeners" of violence-advocating speech consume that speech through their phones and computer screens, and the possible target of that advocated violence might be a long distance from the listeners.[168] Listeners are not likely to respond to such advocacy with immediate action. The angry mob does not rise up from their keyboards when they are incited; they mostly just type back.

Another justification for the punishability of incitement is the lack of opportunity for counterspeech that could minimize the likelihood of the listeners' violent acts. There is no doubt that online anonymity, combined with geographical and temporal dislocation between speaker and audience, has facilitated an increase in the hatefulness of hate speech, its prevalence, and possibly its dangerousness as well.[169] Social scientists have described the reduction in empathy that occurs when interactions that once took place face-to-face and in real time are moved to online and asynchronous settings as the "online disinhibition effect."[170] In short, anonymous internet speech disassociates both the speaker and the object of the speaker's hate from their respective personhoods, and is thus largely consequence-free in terms of social cost.[171]

But the internet has turbocharged the capacity not just for hate speech, but also for counterspeech to that hate speech. One study found that

166. *See* Hess v. Indiana, 414 U.S. 105, 109 (1973) (Rehnquist, J., dissenting) (stating the intent to incite may be inferred from the "import of the language" of the speaker).

167. Healy, *supra* note 145, at 716.

168. As Alexander Tsesis observes:

> Someone surfing the Web can encounter statements that might have led to a fight had they been uttered during the course of a proximate confrontation, but when long distances separate the speaker and intended target it is likely that any pugilistic feelings will dissipate, even if the two happen to meet at some distant point in the future.

Alexander Tsesis, *Inflammatory Speech: Offense Versus Incitement*, 97 MINN. L. REV. 1145, 1173 (2013).

169. *See* Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEX. TECH L. REV. 147, 148–49 (2011); Enrique Armijo, *Meet the New Governors, Same as the Old Governors*, *in* THE PERILOUS PUBLIC SQUARE: STRUCTURAL THREATS TO FREE EXPRESSION TODAY 352, 356–57 (David Pozen ed., 2020).

170. Christopher Terry & Jeff Cain, *The Emerging Issue of Digital Empathy*, AM. J. PHARM. EDUC., May 2016, at 2, https://www.ajpe.org/content/80/4/58 [https://perma.cc/EA6P-483T].

171. *Id.*

hashtagged conversations of controversial topics on Twitter enabled responses to hateful, harmful, or extremist messages that, in some cases, caused the initial user to recant or apologize for their message.[172] Some scholars have argued that the filter bubbling and fake news-enabling associated with social media platforms undermines counterspeech doctrine's applicability with respect to online speech.[173] Others argue that, to the contrary, political communication via social media exposes speakers to differing viewpoints much more often than criticisms of the internet suggest.[174] For purposes of incitement doctrine, however, there is no question that social media platforms create opportunities for counterspeech that are relevant to an imminence analysis.[175] Where counterspeech can occur between advocacy and illegal action, punishable incitement is less likely to be found. Through their design, Twitter, Facebook, and YouTube make space for counterspeech, and thus speech on those platforms is less likely to cause imminent lawless action as that term is understood in the incitement doctrine.[176]

One need not reach too far back into the past for a hypothetical that demonstrates the danger of holding platforms liable for third-party speech alleged to incite violence. In June 1995, *The Washington Post* received in the mail "Industrial Society and Its Future," a 35,000-word manifesto by Ted Kaczynski, known in intelligence circles and the media as the

172. Susan Benesch et al., *Counterspeech on Twitter: A Field Study*, DANGEROUS SPEECH PROJECT 31 (2016), https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/ [https://perma.cc/P63Z-QBJT].

173. *See, e.g.*, Philip M. Napoli, *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble*, 70 FED. COMMC'NS. L.J. 55, 66–67 (2018); SOLOMON MESSING & SEAN J. WESTWOOD, SELECTIVE EXPOSURE IN THE AGE OF SOCIAL MEDIA: ENDORSEMENTS TRUMP PARTISAN SOURCE AFFILIATION WHEN SELECTING NEWS ONLINE 17 (2012), https://cpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/d/2314/files/2021/03/MessingWestwood2014.pdf [https://perma.cc/WQ6Z-W5W4].

174. Frederik J. Zuiderveen Borgesius et al., *Should We Worry About Filter Bubbles?*, 5 INTERNET POL'Y REV., Mar. 2016, at 10, https://policyreview.info/node/401/pdf [https://perma.cc/JU7K-447A].

175. Some academics argue, however, that some platforms that host incitements to violence are intentionally designed to impede or shut out counterspeech. *See, e.g.,* Adrienne Massanari, *#Gamergate and the Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures*, 19 NEW MEDIA & SOC'Y 329, 337 (2017).

176. Incitement is also punishable on the ground that in the absence of its necessary conditions, listeners have time to reflect on the illegal acts advocated by the speaker and decide not to commit them. Healy, *supra* note 145, at 708–09, 717–18. Social media users that come across inciting speech online, almost by definition, have the opportunity to engage in such reflection. *See* Lidsky, *supra* note 169, at 150 (stating that speakers using "social media that permit one-to-many communications . . . are rarely held liable for provoking violence because time for reflection is built into the medium itself"); *see also* Chen, *supra* note 163, at 395 ("Unlike speech spurring on an angry mob, there may be a substantial lag between when speech is posted on a web page or Facebook and when an audience member reads and acts on that speech.").

Unabomber.[177] The correspondence accompanying the manifesto included a threat: if the newspaper published the manifesto, the author would stop harming people.[178] If the newspaper declined, the author would "start building [the] next bomb."[179] Upon receipt of the threat, the *Post*'s leadership reached out to the FBI and DOJ, and on recommendation of Director Louis Freeh and Attorney General Janet Reno, the *Post* published the manifesto in a special section on September 19 of that same year.[180]

Fortunately, the Unabomber did not claim any additional victims after the *Post*'s running of his piece, in large part because its publication assisted the FBI in his capture.[181] But imagine if the Unabomber's threat was reversed—"if you publish, I will kill again"—and he killed another victim after the manifesto was published. Imagine further that the manifesto encouraged a like-minded individual to engage in similar acts, resulting in another death by bombing. There is no interpretation of First Amendment doctrine that would have allowed *The Washington Post* to be held liable for incitement or for aiding-and-abetting murder for its publication of the manifesto in either case. But those who would seek to hold social media companies responsible for failing to take down terrorist speech would seem to have no difficulty finding liability for the platforms—even criminal liability—based on an alleged "dissemination" of the offending speech that is more passive than the *Post*'s affirmative decision to publish the Unabomber's manifesto.[182]

This is not to say, however, that online speech can never be incitement in the First Amendment sense of the term. For example, take a Facebook posting calling on its viewers to "kill a Black person at the Juneteenth parade," or a call to riot at a local mall on an evening later that week, along with an emoji of a gun pointed at a police officer's head:

177. Paul Farhi, *How Publishing a 35,000-word Manifesto Led to the Unabomber*, WASH. POST (Sept. 19, 2015), https://www.washingtonpost.com/lifestyle/style/how-publishing-a-35000-word-manifesto-led-to-the-unabomber/2015/09/18/e55229e0-5cac-11e5-9757-e49273f05f65_story [https://perma.cc/4NVE-5JYX]. The *New York Times* received the manifesto as well, but only the *Post* published it. *Id.*

178. *Id.*

179. *Id.*

180. *Id.*

181. Kaczynski's brother, who recognized Kaczynski's writing style in the published manifesto, alerted law enforcement as to his identity. *Id.*

182. *See* Alexander Tsesis, *Social Media Accountability for Terrorist Propaganda*, 86 FORDHAM L. REV. 605, 616 (2017).

These hypotheticals—the second of which is based on an actual arrest[183]—may turn the corner from abstract advocacy to solicitation of a crime, and the fact that they provide a specific time, place, and/or victim for listeners to commit violent acts might call for a relaxation of the imminence requirement or to place less weight on the capacity-for-counterspeech factor described above. But these are arguments applicable to certain calls to violence generally, not to those made via online speech specifically. And they do not resolve the real issue of focus here: whether incitement-based republication liability for social media platforms for the speech of others can or should exist at all.

183. *Greensboro Teen Arrested, Accused of Using Social Media to Urge a Riot*, MYFOX8.COM (Apr. 28, 2015, 7:17 PM), https://myfox8.com/2015/04/28/greensboro-teen-arrested-for-using-social-media-to-urge-a-riot/ [https://perma.cc/9SYD-SSK4]. The seventeen-year-old youth that posted the above picture to Facebook was charged with violating North Carolina General Statute section 14-288.2 (West 2019), which makes it a misdemeanor to "willfully incite[] or urge[] another to engage in a riot" if either a riot results or "a clear and present danger of a riot is created." The post was made the day after protests and riots in Baltimore stemming from the death of Freddie Gray while in police custody; the photo above right of the emoji image is of those riots, and the photo above left is of the Four Seasons Town Centre, the largest shopping mall in Greensboro.

B. *The "Disinformation" Problem: Fake News as Protected Speech*

In addition to disseminating violence-inciting speech, scholars, policymakers, and journalists have criticized social media platforms for spreading "fake news": fabricated news articles and advertisements based on false information intended to influence voters by use of deceit. The initial social science literature studying sharing and consumption of political information via social media has found, among other things, the following: First, fake news about the 2016 U.S. presidential election was shared faster and more widely than mainstream news stories.[184] Second, enabled by social media microtargeting technology, many fake news ads during the election were aimed at select groups in attempts to suppress or encourage votes and support in that subgroup.[185] Third, "[t]rust in information accessed through social media is lower than trust in traditional news outlets."[186] Fourth, despite claims about "filter bubbles," social media actually increases its users' exposure to a variety of politically diverse news and information relative to traditional media or in-person interaction.[187] Fifth, fake news favored Donald Trump over Hillary Clinton by a wide margin.[188] The fact that Russian government-funded propagandists were behind most of these efforts has led to claims that fake news is a threat to U.S. democracy and the integrity of its elections.[189]

But any government efforts to address the issue of fake news run into First Amendment problems. Recognition of the role that false speech plays in the First Amendment's approach to finding truth runs deep. John Stuart Mill's *On Liberty* recognized that "[f]alse opinions have value . . . because they provoke people to investigate the proposition [at

184. *See* Soroush Vosoughi et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1147 (2018); Craig Silverman, *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook*, BUZZFEED NEWS (Nov. 16, 2016, 5:15 PM), https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook [https://perma.cc/BX6M-J3XK].

185. Abby K. Wood & Ann M. Ravel, *Fool Me Once: Regulating "Fake News" and Other Online Advertising*, 91 S. CAL. L. REV. 1223, 1229 (2018) (citing Indictment ¶ 6, United States v. Internet Rsch. Agency LLC, No. 1: 18-cr-00032 (D.D.C. Feb. 16, 2018), https://www.justice.gov/file/1035477/download/ [https://perma.cc/G486-T5SA]).

186. Hunt Allcot & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*, 31 J. ECON. PERSPS. 211, 212 (2017).

187. Michael A. Beam et al., *Facebook News and (De)Polarization: Reinforcing Spirals in the 2016 US Election*, 21 INFO., COMMC'N. & SOC'Y, Mar. 2018, at 4.

188. Allcot & Gentzkow, *supra* note 186, at 212 ("Our database contains 115 pro-Trump fake stories that were shared on Facebook a total of 30 million times, and 41 pro-Clinton fake stories shared a total of 7.6 million times.").

189. Sabrina Siddiqui, *Half of Americans See Fake News as Bigger Threat Than Terrorism, Study Finds*, GUARDIAN (June 7, 2019, 8:53 AM), https://www.theguardian.com/us-news/2019/jun/06/fake-news-how-misinformation-became-the-new-front-in-us-political-warfare [https://perma.cc/SE5C-D5UP].

issue] further, thereby leading to discovery of the truth."[190] To be sure, some argue that the "collision" of truth and error that Mill described in 1859 does not occur in the social media information ecosystem, where content intended to deceive is easy to produce and free to distribute.[191] Rather than colliding with truth, fake news, turbocharged by bots and by partisans who believe its messages to align with their political beliefs, swallows truth, like the amoeba that swallows the healthy cells in its path. But the Supreme Court has consistently found that false speech deserves First Amendment protection. In *United States v. Alvarez*,[192] the Court held that harmless lies were protected by the First Amendment, and so the Stolen Valor Act, which criminalized falsehoods about military honors, was unconstitutional.[193] Upholding the Act, the Court stated, "would endorse government authority to compile a list of subjects about which false statements are punishable"—a power with "no clear limiting principle."[194] The First Amendment stood in the way, the Court declared, of "the idea that we need Oceania's Ministry of Truth."[195]

Given *Alvarez*'s warnings concerning the crafting of official definitions of "truth" for the purpose of regulating to promote it, it would seem impossible for government to provide a remedy for social media distribution of "disinformation" such as fake news. Even if the government's interests in curbing disinformation in the political speech market or in preventing foreign influence in U.S. elections are compelling, regulating false speech would require government ascertainments of what constitutes the truth—the very concern expressed by the Court in *Alvarez*.[196] In addition, a multitude of alternatives to that process that would be less restrictive of speech are available for that job, including, as also recognized in *Alvarez*, many that do not require government action at all, including counterspeech.[197] It turns out that what Mill said 160 years ago is no less true today, even in the age of social media.

---

190. Daniela C. Manzi, Note, *Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News*, 87 FORDHAM L. REV. 2623, 2626 (2019) (citing JOHN STUART MILL, ON LIBERTY, *reprinted in* ON LIBERTY, UTILITARIANISM AND OTHER ESSAYS 5, 15, 18–54 (Mark Philip & Frederick Rosen eds., 2015)).

191. *See, e.g.*, Tim Wu, Essay, *Is the First Amendment Obsolete?*, 117 MICH. L. REV. 547, 550 (2018).

192. 567 U.S. 709 (2012) (plurality opinion).

193. *Id.* at 709, 715–18.

194. *Id.* at 723.

195. *Id.*

196. And of course, the Supreme Court's most important First Amendment case of all time—*New York Times Co. v. Sullivan*—was a "fake news" case, in that the factual and allegedly defamatory statements at issue in the case were indisputably false. 376 U.S. 254, 271 (1964).

197. *Alvarez*, 567 U.S. at 726.

III. WHAT CAN GOVERNMENT DO?

Though the challenges presented by social media platforms are significant, governments are not powerless to address them. The tools with which to do so are the same ones used in traditional speech markets: measures that all share the goal of providing more information to listeners, not less.

A. *Speaker-Based Disclosures*

Post-*Alvarez*, it seems clear that the government lacks the power to use law to target speech based on its "falsity and nothing more."[198] Such a law or regulation would necessarily be aimed at false statements, and would thus be labeled as content-based, subjected to strict scrutiny review, and certainly found unconstitutional.[199] However, courts have held that speaker-based disclosures—requirements that speakers or their sponsors divulge their identities as a condition of being able to speak—have numerous salutary First Amendment-related benefits, especially in the political speech context.[200] Disclosure-based regulatory models can thus alleviate some of the disinformation-related issues unique to social media.

For example, the Honest Ads Act, which was reintroduced in the U.S. Senate in May 2019, aims to improve the transparency of online political advertisements by imposing several existing disclosure-related laws and regulations to paid internet and digital ads.[201] Consistent with federal law barring foreign campaign contributions, the Act would require platforms, television, and radio stations to "make reasonable efforts to ensure that [electioneering] communications . . . are not purchased by a foreign

198. *Id*. at 719; *see also id.* ("[T]he Court has been careful to instruct that falsity alone may not suffice to bring the speech outside the First Amendment.").

199. *Id*. at 715. The government is not precluded, however, from punishing the most harmful forms of false speech. Courts have found that anti-hoax statutes, which make illegal false reports of emergencies such as terrorist attacks, do not violate the First Amendment. For example, in *United States v. Brahm*, 520 F. Supp. 2d 619, 621–22 (D.N.J. 2007), a poster on the 4chan message board claimed that several "dirty" explosive devices would be detonated at seven NFL games on a specific date. The court found the statute to be valid due to the compelling interest in preserving emergency services for actual threats, and the statute was not overbroad. *Id.* at 628.

200. Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 371 (2010) (noting that compelled disclosure of the identities of those making political expenditures serves the First Amendment by enabling "the electorate to make informed decisions and give proper weight to different speakers and messages"); *see also* Buckley v. Valeo, 424 U.S. 1, 66–67 (1976) (per curiam) (upholding a disclosure requirement in the context of campaign contributions to support the government interests of deterring corruption, alerting voters to the interests to which a candidate is most likely to be responsive, and recordkeeping).

201. Honest Ads Act, H.R. 4077, 115th Cong. § 2 (2017). The Honest Ads Act's reintroduction highlighted Special Counsel Robert Mueller's findings of significant Russian interference in the 2016 presidential election. *See id*. § 3(1).

national, directly or indirectly."[202] It would also impose "public file"-related recordkeeping obligations currently in effect for broadcasters on social media platforms that accept political advertising.[203]

To be sure, political advertising-related disclosures would only cover those stories that use paid content for their dissemination, which covers most, but by no means all, attempts to use social media to deceive or mislead voters. The 2016 election demonstrated that users "happily circulate news with contested content as long as it supports their candidate," regardless of how that content initially showed up in their feed.[204] But limiting the disclosure to advertisements might also cause reviewing courts to apply to the Act the more forgiving intermediate scrutiny standard of review applicable to commercial speech.[205] Also, increasing the availability of information about online advertisers would better assist social media users to assess the validity of those advertisers' messages, even where the messages have been forwarded by a trustworthy source. This would include advertising that, like fake news, is intended to mislead.

B. *Labeling Deep Fakes*

Another challenge associated with modern political speech is that of the "deep fake": a manipulation of existing video and audio through the use of technology and artificial intelligence, usually intended to misrepresent politicians.[206] In May 2019, a video of a speech by U.S. Speaker of the House Nancy Pelosi was slowed down to 75%, which was intended to make Pelosi appear to slur her speech.[207] In response to this and other similar efforts, California passed a law making illegal the distribution of "materially deceptive" audio or visual media with the intent to "injure the candidate's reputation or to deceive a voter into voting for or against the candidate," unless the media is labeled as "manipulated."[208] New York and Texas followed suit, the House Intelligence Committee held a hearing on deep fakes and AI in June

202. *Id*. § 9(c).

203. *Id*. § 8. As the Fourth Circuit recognized in finding a state law similar to the Honest Ads Act violated the First Amendment, burdensome disclosure requirements for political advertising can disincentivize online platforms from carrying political ads altogether. *See* Washington Post v. McManus, 944 F.3d 506 (4th Cir. 2019).

204. Wood & Ravel, *supra* note 185, at 1270–71.

205. *See* Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 566 (1980) (applying a four-part intermediate scrutiny test to a commercial speech case).

206. *See* Lauren Feiner, *Facebook Says the Doctored Nancy Pelosi Video Used to Question Her Mental State and Viewed Millions of Times Will Stay Up*, CNBC (May 24, 2019, 11:31 AM), https://www.cnbc.com/2019/05/24/fake-nancy-pelosi-video-remains-on-facebook-and-twitter.html [https://perma.cc/P6Y2-PT9V].

207. *Id.*

208. CAL. ELEC. CODE § 20010 (West 2020).

2019,[209] and two federal bills, the Malicious Deep Fake Prohibition Act of 2018[210] and the DEEP FAKES Accountability Act,[211] have been introduced in the Senate and House respectively.

Despite the California law's broad application to any manipulated video of a candidate, there is an arguable basis for distinguishing between types of deep fakes. Those manipulations of audio and video that are obviously fake might be better candidates for constitutional protection on the ground they are more akin to "whimsy, humor or satire."[212] As Professor Cass Sunstein writes, "if people do not believe that a deep-fake is real"—if there is no possibility of deception—"there should be no harm."[213]

However, there are compelling governmental interests in minimizing the harm caused by deep fakes, both to the political process generally, which relies on voters' access to truthful information, and to the reputations of those who are depicted in them.[214] Consistent with these interests and cognizant that disclosure is always an alternative less harmful to speech than punishing it outright, the government may be able to mandate that platforms label deep fakes as altered where the platforms are able to do so.[215]

CONCLUSION

In June 2016 and March 2019, Facebook Live brought to the world's attention two unspeakable acts of violence. Seconds after her boyfriend Philando Castile was shot seven times by a police officer during a routine traffic stop in suburban St. Paul, Minnesota, Diamond Reynolds took to Facebook's livestream to narrate the interaction between Castile and the officer that had just occurred, documenting her own arrest, and to show her boyfriend's bloodied body and last gasps.[216] Protests followed, first in Minnesota and then across the country, for nearly two weeks. Nearly two years later, a white supremacist strapped on a GoPro camera and livestreamed himself for seventeen minutes, as he traveled to and entered

209. *House Intelligence Committee Hearing on "Deepfake" Videos*, C-SPAN (June 13, 2019), https://www.c-span.org/video/?461679-1/house-intelligence-committee-hearing-deepfake-videos. [https://perma.cc/KR6M-QVDH].

210. Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (2018).

211. DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019).

212. Cass R. Sunstein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 387, 420 (2020).

213. *Id.*

214. *Id*.

215. Richard L. Hasen, *Deep Fakes, Bots, and Siloed Justices: American Election Law in a "Post-Truth" World*, 64 ST. LOUIS U. L.J. 535, 549 (2020).

216. Pam Louwagie, *Falcon Heights Police Shooting Reverberates Across the Nation*, STAR TRIB. (July 8, 2016, 3:15 PM), http://www.startribune.com/falcon-heights-police-shooting-reverberates-across-the-nation/385861101/ [https://perma.cc/GZS6-S3PD].

the Al Noor Mosque in Christchurch, New Zealand, where he eventually gunned down forty-two Muslims worshiping there.[217]

Though undoubtedly a tragedy, the Castile killing and several similar incidents, including Darnella Frazier's recording and posting of Minneapolis police officer Derek Chauvin's killing of George Floyd, have brought unprecedented transparency and exposure to issues of police shootings in the United States. Social media and livestreaming have empowered Black Lives Matter and other advocates for underrepresented and marginalized communities to raise awareness of the sometimes-deadly realities associated with minorities' interactions with police.[218] Simply put, with the exception of the time period around the Rodney King riots, prior to Facebook and Twitter, police brutality against African Americans was an issue that received little—if any—attention outside of underrepresented communities. In a traditional media-dominated world, those seeking to bring light to the issue could not break through the gatekeepers to reach those in power. Prosecutions of police officers for excessive uses of force are rare to say the least, and those that are brought are mostly unsuccessful. But before social media and smartphones, such cases were barely brought at all.[219] Those same technologies, however, enabled the Christchurch murderer to bring attention to both his act and the ideology that fueled it. So, the question becomes, is the horror of the Christchurch livestreaming the price we pay for the greater knowledge we have gained about police brutality?

It is difficult to craft a liability rule or regulatory regime that would immunize streaming of the Castile shooting aftermath but not the Christchurch massacre. But that is the burden of those who seek to expand potential civil and criminal liability for social media platforms' carriage of third-party speech. The current immunity regime, driven by Section 230 but informed by the First Amendment, permits us to have both. Revising that regime may cause us to have neither.

---

217. Nick Perry & Juliet Williams, *New Zealand Digs Graves as Mosque Massacre Toll Rises to 50*, ASSOCIATED PRESS (Mar. 16, 2019), https://apnews.com/article/massacres-ap-top-news-international-news-christchurch-new-zealand-2aeb7ec2f4d54dad938d3b90089aea79 [https://perma.cc/25LK-W87Y].

218. *See, e.g.*, Christine Hauser et al., *"I Can't Breathe": 4 Minneapolis Officers Fired After Black Man Dies in Custody*, N.Y. TIMES (May 26, 2020), https://www.nytimes.com/2020/05/26/us/minneapolis-police-man-died.html [https://perma.cc/UEE2-NJSR].

219. Tim Nelson et al., *Officer Charged in Castile Shooting*, MINN. PUB. RADIO (Nov. 16, 2016, 4:30 PM), https://www.mprnews.org/story/2016/11/16/officer-charged-in-castile-shooting [https://perma.cc/XRK6-N7XZ].