THE C. BOYDEN GRAY
Center *for the* Study *of the* Administrative State
ANTONIN SCALIA LAW SCHOOL • GEORGE MASON UNIVERSITY

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

Enrique Armijo

CSAS Working Paper 20-10

Should Internet Platform Companies Be Regulated – And If So, How?

GEORGE MASON UNIVERSITY
ANTONIN SCALIA LAW SCHOOL

Gray Center for the Study of the Administrative State
Spring 2020 Public Policy Conference:
Should Internet Platform Companies Be Regulated—And, If So, How?

Conference Working Paper: Reasonableness as Censorship: Algorithmic Content Moderation, The First Amendment, and Section 230 Reform

Enrique Armijo[1]

INTRODUCTION

For the first time in the relatively brief history of the Internet, revising the Communications Decency Act (CDA)'s Section 230 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 functions Section 230 may have once served in "creating the modern Internet,"[2] in the words of one influential critique "[t]oday, huge social networks and search engines enable the rapid spread of destructive abuse."[3] Section 230's immunity from most republisher and distributor-based liability for platforms 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.

Many of these calls are built around the related and longstanding common law liability principles of duty and reasonableness. The use of reasonableness in the Section 230 context would condition the liability of social media platforms, via either "judicial interpretation or legislat[ive]" amendment, on a requirement that the platforms "take[]

[1] Associate Dean for Academic Affairs and Professor, Elon University School of Law and Affiliated Fellow, Yale Law School Information Society Project; Board of Academic Advisors, Free State Foundation. The Gray Center for the Study of the Administrative State at George Mason University supported my participation at the Public Policy Conference on the topic of internet platform regulation, as well as the development of this Working Paper. Thanks as well to Kathryn Romo for outstanding research assistance.

[2] Jeff Kosseff, *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; *see also generally* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

[3] Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity*, 86 FORDHAM L. REV. 401, 411 (2017). *See also* Bobby Chesney & Danielle Citron, *Deep Fakes*: *A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL L. REV. 1753 (2019); Alexander Tsesis, *Terrorist Speech on Social Media*, 70 VAND. L. REV. 651, 654 (2017) ("The internet is awash with calls for terrorism.").

reasonable steps to prevent or address unlawful uses of [their] services."[4] Reliance on duty as a theoretical legal hook for possible intermediary liability in moderating third-party content is taking hold in the United Kingdom as well. 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.[5] The UK government is currently deciding whether the regulator enforcing this duty of care should have the power to block websites and "disrupt business activities" in the event of a platform's breach of the duty.[6] And calls for new regulatory regimes for social media in the United States, with new federal agencies to implement them, advocate for similar approaches.[7]

Other 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."[8] 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."[9] And Hawley's

[4] Citron & Wittes, *supra* note 3, at 419. *See also* Ryan Hagemann, *A Precision Regulation Approach to Stopping Illegal Activities Online*, IBM POLICY BLOG (July 10, 2019), https://www.ibm.com/blogs/policy/cda-230/.

[5] Jeremy Wright & Sajid Javid, *Online Harms White Paper*, DEP'T FOR DIGITAL, CULTURE, MEDIA & SPORT, HOME DEP'T (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf. The government's response to comments on the White Paper ratified this approach, noting that under the regulations to be adopted, stating that as 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." *Online Harms White Paper – Initial Consultation Response*, DEP'T FOR DIGITAL, CULTURE, MEDIA & SPORT, HOME DEP'T (Feb. 12, 2020), https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response.

[6] *See Online Harms White Paper – Initial Consultation Response*, *supra* note 5, at ¶ 57.

[7] *See, e.g.*, 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 (calling for a new federal "Digital Democracy Agency" that would regulate around issues of disinformation, privacy, and promoting local journalism).

[8] "Deplatforming" refers to platforms' removal of users for violations of the platforms' terms of service. *See, e.g.*, Rachel Kraus, , *2018 was the year we (sort of) cleaned up the interne*t, MASHABLE (Dec. 26, 2018), https://mashable.com/article/deplatforming-alex-jones-2018/. "Shadowbanning" refers to platforms' blocking or partially blocking 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).

[9] Ending Support for Internet Censorship Act, S. 1914, 116th Cong. (2019).

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.[10] An analogous bill introduced in the House, the "Stop the Censorship Act," would limit platforms' immunity for blocking content under Section 230 for only content that is "unlawful."[11] There are several more modest legislative efforts at Section 230 reform targeted at specific types of third-party content; one representative proposal is Senator Lindsey Graham's Eliminating Abusive and Rampant Neglect of Internet Technologies (or EARN IT) Act, which would revoke platforms and websites' absolute immunity for distribution of third-party child pornography over their platforms.[12] Under the EARN IT Act, if a company either i) acts consistent with the "best practices regarding the prevention of online child exploitation conduct" developed pursuant to a Commission established by the statute or ii) has "implemented reasonable measures" relating to online child exploitation, it would remain immune.[13]

And these efforts are not limited to the legislative branch or academia. Following these legislators' lead, and in response to complaints about platforms' bias against conservatives, the Trump administration is reportedly considering an executive order tentatively titled "Protecting Americans from Online Censorship" that would, among other things, narrow the Executive Branch's interpretation of Section 230 immunity and require the Federal Communications and Trade Commissions to report on platforms' content moderation practices and whether they are enforced in politically "neutral" ways.[14] Some Democrats seeking to run against President Trump in 2020 have also spoken out against 230 immunity, with former Vice President Joe Biden calling for it to be "revoked, immediately" on the ground Facebook and other platforms are "propagating falsehoods they know to be false." According to Biden, Mark Zuckerberg, Facebook, 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.[15]

[10] *Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*, 1:46:25, C-SPAN (Apr. 10, 2018), https://www.c-span.org/video/?443543-1/facebook-ceo-mark-zuckerberg-testifies-data-protection&start=6378 [hereinafter *Zuckerberg Hearing*]. *See also* Catherine Padhi, *Ted Cruz v. Section 230: Misrepresenting the Communications Decency Act*, LAWFARE (Apr. 20, 2019), https://www.lawfareblog.com/ted-cruz-vs-section-230-misrepresenting-communications-decency-act.

[11] Stop the Censorship Act, H.R. 4027, 116th Cong. (2019).

[12] *See* EARN IT Act of 2020, S.3398, 116th Cong. (2019), available at https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=3165&context=historical.

[13] *Id.*

[14] 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*; 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 (reporting that FCC and FTC officials expressed concerns that such a proposal would violate the First Amendment).

[15] *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; *see also Connecting the Dots: Combating Hate and Violence in America*, *infra* note 81.

This Paper 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 legal 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 below shows the problems associated with adopting a common law-derived standard of civil liability like "reasonableness" as a baseline for prospective intermediary fault. It also discusses the particular challenges that the use of artificial intelligence presents to the task of defining reasonableness, and discusses products liability, another common law theory of fault increasingly being considered as a method for finding platforms liable for third-party content. 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 that attempt to remediate harms caused by third-party content on social media might be possible.

I. COMMON LAW RIGHTS AS REGULATORY WRONGS

A. The "Reasonableness" Problem

The concepts of duty and reasonableness have a long pedigree in the Anglo-American common law of negligence. We owe a duty of care to those whom our conduct might foreseeably injure. The content of that duty of care is said to be defined by reasonableness. When an act or omission causes another physical or another type of harm covered by negligence, the harm-causing party's conduct will be measured by what a reasonable person would have done under the circumstances. In a tort 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. 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 giving rise to liability. If the actor in question's act or failure to act fell below the standard that a plaintiff alleges and a factfinder determines was reasonable, then liability with respect to the harm caused by that act or failure to act is appropriate. Furthermore, past harms define what possible future harms are or should have been foreseeable.

Across a range of domains, the government regularly adopts private tort law-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.[16] In its

[16] Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. (forthcoming) (manuscript at 35), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350508 (citing Daniel J. Solove &

promulgation of new car safety standards, the National Highway Traffic Safety Administration is statutorily required to consider whether a proposed standard is reasonable.[17] 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.[18] 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 standards, in the absence of preemption the plaintiff can point to the noncompliance as evidence of breach of duty in a civil negligence suit. And 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.[19] But the specific dynamic of social media platforms—where the entity to be regulated moderates the expressive content of third parties, and that moderation is the conduct the government intends to regulate under many of the aforementioned reform proposals—fits much less 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 manufacture of products with which others will interact specifically. Such an approach, 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 healthcare—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 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, and Tumblr, and in their online comments sections, the *New York Times* and *Washington Post*. With a few statutory exceptions not discussed in this Paper, all of these companies enjoy

Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 643 (2014)).

[17] 49 U.S.C. § 30111(b)(3) (2012).

[18] David Zaring, *Rule by Reasonableness*, 63 ADMIN. L. REV. 525, 543–45 (2011) (citing, *inter alia¸* the Securities Act of 1933 § 11, 15 U.S.C. § 77k (2012), FIN. INDUSTRY REG. AUTHORITY, FINRA MANUAL RULE 2310, and several provisions of the Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, 12 U.S.C. § 5322 (2012), all of which apply a reasonableness standard to a range of conduct and enforcement actions in the financial sector).

[19] RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. LAW INST. 1979); *id.* § 12 (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").

immunity for third-party content under Section 230, 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 that content. And 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 available capital and technological sophistication, to adopt more holistic and responsive content moderation regimes, including those that use artificial intelligence (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.[20] The result of adopting a reasonableness standard will thus very likely be the very state of affairs that many of those advocating for the change most want to avoid—an entrenchment of the largest social media companies as hosts of third-party speech, an increase in their power over 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.[21] It is easy to see how such a technology might also be helpful for the company to intervene if passengers are placed in danger by drivers. When a passenger is injured and makes a claim that Uber's failure to use geolocation to avoid harm to the passenger was a cause of the passenger's 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, 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 ride-sharing services more generally with respect 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 developed and used by Uber—particularly given how susceptible juries are to hindsight-related biases and heuristics, which cause

[20] Eric Goldman, *Want to Kill Facebook and Google? Preserving Section 230 is Your Best Hope,* BALKINIZATION.COM (June 3, 2019, 9:30 AM), https://balkin.blogspot.com/2019/06/want-to-kill-facebook-and-google.html (saying 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").

[21] *See, e.g.*, Uber Engineering Blog, *How Uber Engineering Increases Safe Driving with Telematics* (June 29, 2016) https://eng.uber.com/telematics/; 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 (discussing Uber's use of telematics technology to track driver safety).

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.[22] Additionally, 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 is 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 and foreseeability, would be *de facto* strict liability for platforms' facilitation of harmful third-party speech.[23]

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 primarily the cause of the complained-of harm. Deciding how to design and manufacture a car or a drug is within the manufacturer's control. The foreseeable harms associated with a particular design, manufacturing process, or warning can be designed around 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,[24] and sometimes a product manufacturer can be held partially liable for harms caused by foreseeable misuses of their products by third parties. 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. 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 with respect to a duty that the *owner owed to the harmed party that was foreseeably on the owner's premises*. 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—but the entire world. The duty to act reasonably does not extend that far.

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*—an act or omission that a reasonable person would undertake to reduce or eliminate a foreseeable risk of harm that the relevant party did not—from *nonfeasance*—a party's failure to act to protect one from a risk of harm caused by another, which in the absence of some other duty-creating doctrine, cannot constitute

[22] *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.") (citing studies and articles).
[23] Montgomery, *supra* note 28, at ___.
[24] RESTATEMENT (SECOND) OF TORTS § 876 (AM. LAW INST. 1979).

unreasonable conduct.[25] As Professor Benjamin Zipursky notes, a claim that a host or other platform has failed to take down the allegedly harmful speech of another party sounds more as nonfeasance than misfeasance[26] (assuming the nonfeasance/misfeasance distinction applies to online defamation at all; Zipursky also argues that the broad judicial interpretations of Section 230's immunity have blocked the common law from meaningfully reaching that question[27]). The nearly 50-year-old *Scott v. Hull* 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. [28] There, the court found that even though the plaintiff had given the defendant landowner notice of the defamatory statement that was graffitied on and visible to the general public from their wall, the building owner could not be held liable as a common law republisher because a failure to take the statement down was mere nonfeasance.[29] Failing to remove "the graffiti merely . . . after its existence was called to their attention," held the court, was not enough of a "positive act" to meet the publication requirement for common law defamation.[30] To characterize a social media platform's failure to take down third-party content as unreasonable thus contravenes the misfeasance/nonfeasance distinction.

Even in cases where more than one party's conduct combines to cause a harm and the degree to which each is responsible is allocated via the comparative fault system, each of those causes are direct, not facilitative, as is the case for prospective intermediary liability for third-party conduct. Defamation actions, for example, generally do not allocate fault as between the speaker and republisher of the defamatory statement at issue. The republisher's affirmative decision to disseminate the defamatory statement—again, under the common law an act of misfeasance, not of nonfeasance[31]—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.[32] It is certainly so that the third party's defamatory or other harmful statement's *reach* is 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, not the predicate question of

[25] 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* DAN B. DOBBS, THE LAW OF TORTS §§ 314–30, at 853–94 (3d ed. 2004).

[26] Benjamin C. Zipursky, *The Monsanto Lecture: Online Defamation, Legal Concepts, and the Good Samaritan*, 51 VAL. U.L. REV. 1, 19–21 (2016).

[27] [Ben cmts at DOJ Sec 230 workshop – find tr]

[28] 259 N.E.2d 160 (Ohio App. 1970).

[29] *Id.* at 161-62.

[30] *Id.* at 162. The *Hull* court distinguished *Hellar v. Bianco*, , 244 P.2d 757 (Cal. App. 1952), an earlier case from California that applied the common law republication rule to a tavern owner who failed to remove a defamatory statement about the plaintiff from their bathroom wall. There, the tavern owner's affirmative act of continuing to holding open of the tavern to invitees who could see the statement that the owner refused to remove was a positive act that both constituted misfeasance and operated as a ratification of the defamation, such that the owner could be as directly liable as the graffitiing original defamer.

[31] Zipursky, *supra* note 26, at 19.

[32] *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).

liability for the harm as measured by those damages, since the publication element of defamation is met by the statement's utterance to just one person other than the plaintiff.[33]

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.[34] Since the economic benefit of any single piece of user-generated content is *de minimis* and potential liability as a result of that content is significant, incentives weigh heavily toward removing content that is even arguably objectionable.[35] This would result in a significantly degraded environment for speech, and, to repeat the point, a huge increase in what many Section 230 reformers consider the greater evil—censorship of platform users' First Amendment-protected speech.

B. The AI Problem

In addition to the general reasonableness-based problems as a basis for content moderation-derived liability described above, any new reasonableness-based standard for intermediary liability would have to take increasing account of the largest platforms' intent to rely more on artificial intelligence 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.[36] With four petabytes of data's worth of postings on Facebook per day, human

[33] RESTATEMENT (SECOND) OF TORTS § 558(b) (AM. LAW INST. 1979).

[34] Daphne Keller, *Who Do You Sue?: State and Platform Power Over Online Speech*, Aegis Ser. Paper No. 1902 (Jan. 29, 2019), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf (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.").

[35] *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 Milligan, *Technological Intermediaries and the Freedom of the Press*, 66 SMU L. REV. 157, 167, 172 (2013) (same). Indeed, the very 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.").

[36] *Zuckerberg Hearing*, *supra* note 10, 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. In later statements, Facebook has hedged its confidence in AI's ability to solve its most difficult content moderation problems. *See, e.g.*, Monika Bickert, *European Court Ruling Raises Questions about Policing Speech*, FACEBOOK NEWSROOM BLOG (Oct. 14, 2019),

review of potentially Standards-offending third-party content will never scale in a way that would satisfy Facebook's users, prospective regulators, and other constituencies. Same with YouTube—it is impossible to prescreen five hundred hours of new video per minute.[37] Given this challenge, Zuckerberg discussed AI as not simply an *ex post* tool that would permit human content moderators to identify Standards-infringing content more quickly, 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. And the current regulatory appetite for greater intermediary liability internationally implicitly relies on the perceived feasibility of a move from *ex post*, 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 increasing use of AI.[38]

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. This work has caused the moderators PTSD-like traumas and drug use, among other stress-related effects.[39] Zuckerberg apparently sees

https://newsroom.fb.com/news/2019/10/european-court-ruling-raises-questions-about-policing-speech/:

> 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.

[37] *See* 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").

[38] Hannah Bloch-Wehba, *Automation in Moderation*, CORNELL INT'L L. J. (forthcoming 2020), (manuscript at 28–34), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619.

[39] *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; 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; SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019); Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, THE VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

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' problem.

As an initial matter, however, we should be skeptical of AI's ability to play a material role in content moderation, particularly context-specific content like defamation or hate speech, with the confidence that Mark Zuckerberg communicated to Congress in 2018.[40] As a general rule, AI "may work better for images than text," and in areas where "there is a consensus about what constitutes a rule violation."[41] The company admitted shortly after Zuckerberg's testimony that its current AI tools only captured about 38 percent of the content that it deemed hate speech in the first quarter of that year.[42] And Twitter engineers recently revealed that algorithms intended to preemptively identify and take down white supremacist-posted material would also sweep up tweets from Republican politicians or their supporters.[43] But putting aside technical feasibility, for present purposes 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, utilizing 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

[40] *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 ("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* (Mar. 2017), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf.

[41] 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 (Houghton Mifflin Harcourt 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 for "text-based profanity, obscenity, or racial slurs" is also effective, but not for flagging more nuanced and context-based content; this 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 37, at 98–100 (describing word filtering moderation processes).

[42] Facebook Newsroom, *Facebook Publishes Enforcement Numbers for the First Time*, FACEBOOK (May 15, 2018), https://newsroom.fb.com/news/2018/05/enforcement-numbers/.

[43] 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.

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 might find an opportunity 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. AI, the argument goes, performs a function; it does not communicate. 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 for restraining the power of platforms and securing fundamental freedoms" for users online.[44]

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.[45] The First Amendment protects human speakers and authors, not machines. But even though the product of most algorithmic authorship is automation, all algorithms begin with human authors.[46] Even automated content moderation is simply a form of editing—"deciding [which content] to publish, withdraw, postpone or alter"[47]—a category of speech that receives full First Amendment protection. 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.[48] A content-moderating algorithm, then, is just expressing the message of the individuals who wrote the code that directs the algorithm to moderate; here, the expressive content of the algorithm's decisions are interpretations and implementations of the platforms' First Amendment-protected terms of service.[49] The content-moderating AI that Mark Zuckerberg envisions for Facebook's future would replicate the 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 the argument that embedded within AI are the biases and value judgments of the AI's creators, often with deleterious effects when those algorithms

[44] Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, LEWIS & CLARK L. REV. (forthcoming) (manuscript at 8–9), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3439261

[45] Selbst, *supra* note 17, at 4.

[46] *See, e.g.*, Stuart Benjamin, *Algorithms as Speech*, 161 U. PA. L. REV. 1445, 1479 (2013) ("[T]he fact that an algorithm is involved does not mean that a machine is doing the talking").

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

[48] Mark Zuckerberg, A Blueprint for Content Governance and Enforcement, FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[49] 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 Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law*, 51 U.C. DAVIS L. REV. 589, 606 (2017).

are applied to members of communities that have been the object of those human-based biases and value judgments.[50] This literature necessarily relies on the presumption that algorithms are speech, since bias (even implicit bias) is expressive in nature. The First Amendment protects substantive communications such as content moderation decisions and their implementation, even if those communications are expressed through the use of artificial intelligence.

So, government officials will not be able to escape First Amendment scrutiny of any efforts to regulate content moderation practices on the ground the moderation is automated via artificial intelligence. Modifying or doing away with 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.

C. The Products Liability Problem

Generally, when an automated process causes a harm, the legal theory supporting compensation for the harmed party is one of strict liability—*i.e.*, liability without fault.[51] Strict products liability theory, like reasonableness, is making inroads for use as a liability theory against online platforms. The platforms' design, so go these arguments, are (1) inherently defective with respect to how they organize, post, or moderate third-party content and other information; (2) those defects caused 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 platform.[52] 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 or difficult to find for purposes of direct suit.[53] Attorneys bringing claims against platforms have increasingly embraced the theory

[50] *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Thomas Davidson et al., Racial Bias in Hate Speech and Abusive Language Detection Datasets (2019), https://arxiv.org/pdf/1905.12516.pdf; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. (forthcoming 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528; Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection* (2019), https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf.

[51] 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: PRODUCTS LIABILITY § 2 (AM. LAW. INST. 1998).

[52] *See, e.g.*, text accompanying notes 37–39; Ari Ezra Waldman, *The Marketplace of Fake News*, 20 U. PENN. J. OF CONST'L L. 845, 848 n.16 (2018) (describing scholarly project that will "show the spread of fake news is a designed-in aspect of online social network platforms. Therefore, I argue that the common law of products liability for design defects offers lawyers and legal scholars several principles for structuring a legal response to fake news.").

[53] *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* 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) (empirical study of § 230 case law finding "41.2% of the decisions studied involved anonymous content.").

as a work-around Section 230.[54] 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 increased skepticism toward Section 230 immunity, 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 vs. Amazon.Com, Inc.*,[55] a woman who purchased a retractable dog leash from a third-party vendor on Amazon.com sued Amazon for strict products liability when she was harmed by a defect in the leash.[56] A divided panel of 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.[57] The court found 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."[58] 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.[59] The Third Circuit has agreed to rehear the case en banc, 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, but misrepresentations in the "marketing" of those products could in theory form a basis for strict intermediary liability under a products-based theory.[60]

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 (or "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," so as to prevent "users to set a time limit that blocks the user's own access to those platforms across all devices," and to "provid[e the] user with an award for engaging with the social media platform"—*i.e.*, badges—that do not "substantially increase access to now or additional services, content, or functionality" on the platform.[61] The Act justifies such an intervention on the ground that

[54] *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.
[55] Oberdorf v. Amazon.com Inc., 930 F.3d 136 (3d Cir. 2019), *rev'd en banc*, 2019 WL 3979586 (3d Cir. Aug. 23, 2019).
[56] *Id.* at 140.
[57] *Id.* at 147.
[58] *Id.* at 153.
[59] *Id.*
[60] *See, e.g.*, *State Farm Fire & Cas. Co. v. Amazon.com, Inc.*, 390 F. Supp. 3d 964, 967 (W.D. Wis. 2019).
[61] SMART Act, S. 2314, 116th Cong. (2019).

"internet companies *design* their platforms and services to exploit brain physiology and human psychology."[62] 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[63]; social media, like cigarettes, is unreasonably and dangerously addictive. The legislation, in other words, regulates social media platform's design because of the harms that design exposes their 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*, 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.[64] 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 app's inability to detect abusive accounts, and its failure to warn its users about abusive uses of the type he was subjected to, was a cause of his harm.[65] But the district court in which the claim against Grindr was filed 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.[66] 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," platforms could not be held liable. In other words, any potential duty to warn a user concerning third-party content is precluded by Section 230. The Second Circuit upheld the district court, and the Supreme Court declined to review the Second Circuit's decision.

[62] *Id.* § 1 (emphasis added).

[63] *See, e.g.*, Evans v. Lorillard Tobacco Co., 990 N.E. 2d 997,1020-21 (Mass. 2013).

[64] 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*, THE OUTLINE (Jan. 11, 2019, 2:02 PM), https://theoutline.com/post/6968/grindr-lawsuit-matthew-herrick?zd=2&zi=mzgo5han.

[65] *See* First Amended Complaint, ¶¶ 108–120, at 26-27, Herrick v. Grindr, LLC, 306 F. Supp. 3d 579 (S.D.N.Y. 2018), *aff'd*, 765 F. App'x 586 (2d Cir. 2019) (No. 1:17-CV-00932) (alleging products liability-based manufacturing and warning defect claims).

[66] Herrick v. Grindr, LLC, 306 F. Supp. 3d 579, 588 (S.D.N.Y. 2018), *aff'd*, 765 F. App'x 586 (2d Cir. 2019). In affirming the District Court's decision, the Second Circuit agreed with this distinction. Herrick v. Grindr, LLC, F. App'x 586, 590 (2d Cir. 2019) (concluding that Herrick's products liability claims were "based on information provided by another information content provider and therefore" were barred by Section 230). The Supreme Court declined Herrick's request to review the Second Circuit's decision. Herrick v. Grindr, LLC, 765 F. App'x 586 (2d Cir. 2019), *cert. denied*, 140 S. Ct. 221 (2019).

Additionally, using strict products liability as a way to find republication liability for social media dissemination of harmful third-party content runs afoul of a different principle, embedded in the First Amendment rather than Section 230: the requirement of scienter, or knowledge of one's own wrongdoing. In 1959's *Smith v. California*, the 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."[67] Strict liability could not be the basis for liability for carrying another's speech, the Court found, 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."[68] 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."[69] 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 *Smith*, 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.[70] As to the latter, where a "plaintiff'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."[71] 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.

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 were to warm to such approaches, they will eventually run afoul of the First Amendment's scienter-related principles as set out in *Smith*.

II. CONSTITUTIONAL PROBLEMS IN A POST-230 IMMUNITY WORLD

[67] 361 U.S. 147, 149 (1959).

[68] *Id.* at 152.

[69] *Id.* at 153.

[70] RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 19, cmt. d (AM. LAW. INST. 1998).

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

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

To begin, an obvious point bears reemphasis, especially given the current regulatory appetite: *content moderation policies are protected speech.* Private parties have brought dozens of cases against Internet platforms complaining of the platforms' decisions to take down their content. In the post-takedown context of civil litigation against individual platforms, courts have unanimously found that content moderation decisions are protected speech by private parties. Despite current debates around regulating social media, there is no reason to assume that increased regulation of content moderation policies would compel a different result.

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.[72] 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."[73] But the Court has also held that the First Amendment protects most *advocacy* of illegal action, with one exception: advocacy that is directed to incite or produce imminent lawless action that is likely to do so.[74] There is a significant amount of speech published on social 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 that speech, or regulating platforms that carry the speech of those speakers. One is the specific intent requirement for

[72] Brown v. Hartlage, 456 U.S. 45 (1982).

[73] U.S. v. Freeman, 761 F.2d 549 (9th Cir. 1985) (opinion by then-Judge Kennedy).

[74] 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.

District of Columbia v. Garcia, 335 A.2d 217, 224 (D.C. App. 1975). *See also* Marc Rohr, *The Grand Illusion?: The* Brandenburg *Test and Speech that Encourages or Facilitates Criminal Acts*, 38 WILLAMETTE L. REV. 1, 27 (2002) (citing *Garcia*); Thomas Healy, *Brandenburg in a Time of Terror*, 84 NOTRE DAME L. REV. 655, 672 (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.").

inchoate crimes like incitement, and the other is the constitutional requirement of imminence.

The primary impediment to regulating platforms' carriage of hate speech advocating violence or other criminal activity is the fifty-year-old *Brandenburg v. Ohio*.[75] 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."[76] 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."[77] 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 to mean near-immediacy.[78]

The reason the common law imposed an imminence requirement was because assault as an avenue for civil liability was directly "tied to failed battery cases"—*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 they were aware of the defendant's attempt.[79] To put it more colloquially, a puncher with bad aim should not escape tort liability because they swung and missed. 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; as Chafee said, the First Amendment permits punishing a speaker for "political 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."[80]

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

[75] Brandenburg v. Ohio, 395 U.S. 444 (1969).

[76] *Id.* at 447 (emphasis added).

[77] RESTATEMENT (SECOND) OF TORTS § 29(1), cmt. b (AM. LAW 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. LAW INST., Tentative Draft No. 1, 2015) (defining "imminent" to mean that "the contact will occur without significant delay"); DOBBS, *supra* note 25, § 34 at 65 (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.").

[78] *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).

[79] RESTATEMENT (THIRD) OF TORTS: INTENTIONAL TORTS TO PERSONS § 105, cmt. e (AM. LAW INST., Tentative Draft No. 1, 2015).

[80] Zechariah Chafee, *Freedom of Speech in War Time*, 32 HARV. L. REV. 932, 963–64 (1919); *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 bad words for their [mere] bad tendency impossible").

speech that advocates violence or other illegal acts.[81] But there are serious problems associated with holding a republisher of incitement liable to the same degree as the initial speaker in the same way as the common law holds the republisher of a defamatory statement equally liable.[82] In *Brandenburg v. Ohio* itself, the government became aware of Clarence Brandenburg's speech after the KKK rally at which he spoke was broadcast as part of a Cincinnati television station's report.[83] There is no indication that the prosecutors considered bringing charges against the station for airing Brandenburg's call to violence along with its 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. 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.[84] In the latter case, the radio station itself was calling for and facilitating the systemic murder of the country's Tutsi population. 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 the last Part, 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 for the crimes the speech facilitated was justified.

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 modification of that content make the possibility of intermediary liability for incitement a closer case. Traditional media often report 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. Unlike defamation, which can give rise to liability based on reckless disregard or even negligence in the case of a private person, a social media

[81] *See, e.g.*, *Connecting the Dots: Combating Hate and Violence in America*, BETO FOR AMERICA, https://www.courthousenews.com/wp-content/uploads/2019/08/beto-gun-plan.pdf ("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*, THE 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.

[82] *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/; 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. PENN. J. OF L. & PUB. AFF. 56 (forthcoming) (manuscript at 24), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330195.

[83] Brandenburg v. Ohio, 395 U.S. 444, 445 (1969).

[84] *See, e.g.*, Eric Paulsen, *Facebook Waking Up to Genocide in Myanmar*, THE DIPLOMAT (Sept. 21, 2018), https://thediplomat.com/2018/09/facebook-waking-up-to-genocide-in-myanmar/; 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/.

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.

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 illegal 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.[85] Punishing only speech likely to incite imminent unlawful activity is also justified from efficiency and deterrence perspectives. As Thomas Healy writes,

> [w]here 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 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.[86]

Incitement, in other words, and like other inchoate crimes, is "designed to interdict a harmful chain of causation once a substantial step has been taken towards commission.[87] The primary motivation for criminalizing inciting speech is to prevent the crimes that speakers would otherwise encourage from being committed in the first place.

This fundamental dynamic changes, however, when the speech 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 listeners to respond to its call for violence or other illegal acts. "[T]he identity of the listeners and the speaker, the place and the crime being advocated,"[88] as well as 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. So as a general matter, incitement-based liability or regulation is difficult to justify when the "listeners" of violence-advocating speech consume that speech from off their phones and computer screens, and the possible target of that advocated violence might be a long

[85] *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, not the listener's.

[86] Healy, *supra* note 74, at 716.

[87] Wilson & Kiper, *supra* note 82, at 82.

[88] Healy, *supra* note 74, at 716.

distance from the listeners.[89] 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; mostly they 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 aggravated an increase in the *hatefulness* of hate speech, and possibly its dangerousness as well.[90] 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"—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.[91]

But the Internet has turbocharged the capacity not just for hate speech, but also for counterspeech to that hate speech. One study found that hashtagged conversations of controversial topics on Twitter permitted responses to hateful, harmful or extremist messages that, in some cases, caused the initial user to recant or apologize for their message.[92] 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.[93] 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.[94] For purposes of incitement doctrine, however, there is no question that social media platforms create opportunities for counterspeech that are relevant to

[89] 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.

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

[90] *See* Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEX. TECH L. REV. 147, 148–49 (2011).

[91] Christopher Terry & Jeff Cain, The Emerging Issue of Digital Empathy, AM. J. OF PHARMACEUTICAL EDUC., May 2016, at 1, https://www.ajpe.org/content/ajpe/80/4/58.full.pdf.

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

[93] *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. COMM. L.J. 55 (2018); SOLOMON MESSING & SEAN J. WESTWOOD, SELECTIVE EXPOSURE IN THE AGE OF SOCIAL MEDIA: ENDORSEMENTS TRUMP PARTISAN SOURCE AFFILIATION WHEN SELECTING NEWS ONLINE (2012), http://www.dartmouth.edu/~seanjwestwood/papers/CRsocialNews.pdf.

[94] Frederik J. Zuiderveen Borgesius et al., *Should we Worry About Filter Bubbles?*, INTERNET POL'Y REV., Mar. 2016, at 1, https://policyreview.info/node/401/pdf.

imminence analysis.[95] Where counterspeech can occur between advocacy and illegal action, punishable incitement is less likely to be found. Twitter, Facebook, and YouTube make space for counterspeech, and thus speech on those platforms is less likely to cause *imminent* lawless action.[96]

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 Unabomber. The correspondence accompanying the manifesto included a threat: if the newspaper published the manifesto, the author would stop harming people. If it declined, he would "start building [the] next bomb."[97] 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.

Fortunately, the Unabomber did not claim any additional victims after the *Post*'s publication of his piece, in large part because its publication assisted the FBI in his capture.[98] But imagine if it did not and the Unabomber killed another victim after the manifesto was published, and imagine further that the manifesto encouraged a like-minded individual to do engage in similar acts, resulting in a 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 either 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

[95] 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(3) NEW MEDIA & SOC'Y 329 (2017).

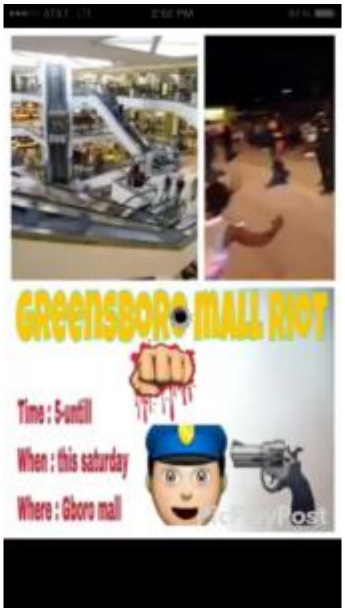
[96] Incitement is also punishable on the ground that in the absence of its necessary conditions, listeners have time to reflect on the illegal acts advocated by the speaker and decide not to commit them. Healy, *supra* note 74, at 708–09, 717–18. Social media users that come across inciting speech online, almost by definition, have the opportunity to engage in such reflection. *See* Lidsky, *supra* note 90, 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 85, 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.").

[97] 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.html. The *New York Times* received the manifesto as well, but only the *Post* published it. *Id.*

[98] Kaczynski's brother recognized his writing style in the published manifesto, which assisted in his capture. *Id.*

"dissemination" of the offending speech that is more passive than the *Post*'s affirmative decision to publish the Unabomber's manifesto.

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:



These hypotheticals—the second of which is based on an actual arrest[99]—may turn the corner from abstract advocacy to solicitation of a crime, and the fact 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.

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

[99] *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/. The 17-year-old youth that posted the above picture to Facebook was charged with violating N.C. Gen. Stat. § 14-288.2 (West, Westlaw through S.L. 2019 Reg. Sess.), 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.

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 and 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, that 1) fake news about the 2016 U.S. presidential election was shared faster and more widely than mainstream news stories[100]; 2) enabled by social media microtargeting technology, many fake news ads ran during the election were aimed at select groups in attempts to suppress or encourage votes and support in that subgroup[101]; 3) "[t]rust in information accessed through social media is lower than trust in traditional news outlets"[102]; 4) 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[103]; and 5) fake news favored Donald Trump over Hillary Clinton by a wide margin.[104] 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.[105]

But any government efforts to address the issue of fake news run into First Amendment problems. The role that false speech plays in the 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 issue] further, thereby leading to discovery of the truth."[106] To be sure, some argue that the "collision" of truth and error that Mill described in 1859 does not occur in the social media

[100] Soroush Vosoughi et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1147 (2018). *See also* 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-newsoutperformed-real-news-on-facebook.

[101] 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 Research Agency LLC, No. 1: 18-cr-00032-DLF (D.D.C. Feb. 16, 2018), https://www.justice.gov/file/1035477/download).

[102] Hunt Allcot & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*, 31 J. OF ECON. PERSPECTIVES 211, 212 (2017).

[103] Michael A. Beam et al., *Facebook News and (de)Polarization: Reinforcing Spirals in the 2016 US Election*, 21 INFO., COMM., AND SOC'Y 4 (2018), https://www.researchgate.net/publication/323565916_Facebook_news_and_depolarization_reinforcing_spirals_in_the_2016_US_election.

[104] Allcot & Gentzkow, *supra* note 101, 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.").

[105] Sabrina Siddiqui, *Half of Americans see fake news as bigger threat than terrorism, study finds*, THE 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.

[106] 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 *On Liberty*).

information ecosystem, where content intended to deceive is easy to produce and free to distribute[107]; 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*, the Court held that "harmless lies" were protected by the Amendment, and so the Stolen Valor Act, which criminalized falsehoods about military honors, was unconstitutional.[108] 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."[109] The First Amendment stood in the way, the Court declared, of "the idea that we need Oceania's Ministry of Truth."[110]

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.*[111] 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.[112] It turns out that what Mill said 160 years ago is no less true today, even in the age of social media.

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 government lacks the power to use law to target speech based on its "falsity and nothing more."[113] Such a law or regulation would necessarily be aimed at false statements, and would thus be content-based, subjected to

[107] *See, e.g.*, Tim Wu, *Is the First Amendment Obsolete?*, 117 MICH. L. REV. 547, 550 (2018).
[108] 567 U.S. 709 (2012).
[109] *Id.* at 723 (plurality).
[110] *Id.*
[111] And of course, the Supreme Court's most important First Amendment case of all time—*New York Times v. Sullivan*—was a "fake news" case, in that the factual and allegedly defamatory statements at issue in the case were false. 376 U.S. 254, 271 (1964).
[112] *Alvarez,* 567 U.S. at 726.
[113] *Id.* at 709; *see also id.* at 719 ("[T]he Court has been careful to instruct that falsity alone may not suffice to bring [] speech outside the First Amendment").

strict scrutiny review, and certainly found unconstitutional.[114] However, courts have held that speaker-based disclosures—*i.e.*, requirements that speakers or their sponsors divulge their identities as a condition of being able to speak—have numerous salutary First Amendment-related benefits.[115] 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.[116] 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 nation, directly or indirectly."[117] It would also impose "public file"-related recordkeeping obligations currently in effect for broadcasters on social media platforms that accept political advertising.[118]

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.[119] 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. 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 much fake news, is intended to mislead.

B. Labeling Deep Fakes

Another challenge associated with modern political speech is that of the "deep fake": manipulations of existing videos and audio through the use of technology and artificial

[114] [cite] 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 message board 4chan 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.

[115] Citizens United v. FEC, 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).

[116] Honest Ads Act, , H.R. 4077, 116th Cong. (2017). The Honest Ads Act's reintroduction highlighted Special Counsel Robert Muller's findings of significant Russian interference in the 2016 presidential election. *See id.* § 3(1).

[117] *Id.* § 9(c)(3).

[118] *Id.* § 8.

[119] Wood & Ravel, *supra* note 100, at 1270–71.

intelligence, usually intended to misrepresent politicians. In May 2019, a video of a speech by U.S. Speaker of the House Nancy Pelosi was slowed down to 75 percent, which was intended to make Pelosi appear to slur her speech. In response to this and other similar efforts, California has passed a law making illegal the distribution of "materially deceptive" audio or visual media with the intent to "injure [a] candidate's reputation or to deceive a voter into voting for or against the candidate," unless the media is labeled as "manipulated."[120] New York and Texas have followed suit, the House Intelligence Committee held a hearing on deepfakes and AI in June 2019,[121] and two federal bills, the Malicious Deep Fake Prohibition Act[122] and the DEEPFAKES Accountability Act,[123] 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"; as Cass Sunstein writes, "if people do not believe that a deep-fake is real"—*i.e.*, if there is no possibility of deception—"there should be no harm."[124]

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.[125] 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.[126]

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, document her own arrest, and show her boyfriend's

[120] Cal. Elec. Code § 20010 (West, Westlaw through Ch. 1 of 2020 Reg. Sess.).

[121] *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.

[122] Malicious Deep Fake Prohibition Act, S. 3805, 115th Cong. (2018).

[123] DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019).

[124] Cass Sunstein, *Falsehoods and the First Amendment* (forthcoming) (manuscript at 23), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426765.

[125] *Id.* at 21–25.

[126] Richard Hasen, *Deep Fakes, Bots, and Siloed Justices: American Election Law in a Post-Truth World*, ST. LOUIS UNIV. L. REV. (forthcoming 2019) (manuscript at 14), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418427.

bloodied body and last gasps.[127] 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 17 minutes, as he traveled to and entered the Al Noor Mosque in Christchurch, New Zealand, where he eventually gunned down 42 Muslims worshiping there.

Though undoubtedly a tragedy, the Castile killing and several other similar incidents 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. Though most manslaughter prosecutions of police officers remain unsuccessful, before social media and smartphones such cases were barely brought at all.[128] Simply put, prior to Facebook and Twitter, police brutality against African Americans was a fringe issue that received little if any attention outside the United States in the 25 years following the Rodney King riots. In a traditional media-dominated world, those seeking to bring light to the issue could not break through the gatekeepers. 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 permit 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 would cause us to have neither.

[127] Pam Louwagie, *Falcon Heights Police Shooting Reverberates Across the Nation*, MINN. STAR-TRIBUNE (July 8, 2016, 3:15 PM), http://www.startribune.com/falcon-heights-police-shooting-reverberates-across-the-nation/385861101/.

[128] Tim Nelson et al., *Officer Charged in Castile Shooting,* MINN. PUB. RADIO (Nov. 16, 2016), https://www.mprnews.org/story/2016/11/16/officer-charged-in-castile-shooting.