University of the Pacific Law Review

Volume 55
Issue 2 *Symposium Free Speech Forum*

Article 11

2-20-2024

AI Accountability & Social-Technical Solutions for the Regulation of Internet Free Speech

Michael Martin Losavio
University of Louisville

AI Accountability & Social-Technical Solutions for the Regulation of Internet Free Speech

Michael Martin Losavio[*]

TABLE OF CONTENTS

Profits and revenue incentivize social media to promote content that is undesirable, or even damaging, to some, as part of algorithmic design to drive and maximize user engagement and advertising revenue. We suggest that social-technical and technical-legal solutions, such as curation algorithms and artificial intelligence systems, may optimize limitations on undesirable content while minimizing the censorship of legitimate speech. These solutions may also support

---

* Associate Professor, Department of Criminal Justice, Instructor, Department of Computer Science and Engineering, University of Louisville, USA. Thanks to Professor Russell Weaver, Professor Leslie Jacobs and the 2023 Luxembourg Free Speech Forum for aiding in the development of these thoughts on technology and free speech in the modern world.

the design of systems where, in Louis Brandeis's formulation, the proper response to bad speech is more and better speech. But Justice Brandeis's formulation came during a time where most new items in mass media were edited and curated under standards of accuracy and fairness, no matter how thin. The law must acknowledge this has changed with wide-open access to media for everyone. And it must address the potential "error rate" in contemporary social media systems and accommodate efforts to implement and improve their reduction of error and the accuracy of social media.

I. INTRODUCTION: SPEECH ON THE INTERNET HAS ALWAYS BEEN REGULATED

The discussion, ferment, and outrage over internet social media content has swiftly moved to assertions of needed accountability for results from their operations. In light of the safe harbor provisions that shield social media operations from liability for user-generated content, this discussion may herald a possible return to accountability for conduct to which everyone else is subject. Under the laws of the United States, social media benefits from statutory immunity for third party content posted on their platforms. This immunity is set by Section 230 of the Communications Decency Act (1996) (CDA)[1]. Ironically, the US firearms industry is one of the few other industrial and commercial domains that has similar immunity for manufacturers and vendors for injuries caused by their activities and products.[2] Each in its own way offers important benefits to their law-abiding users, while opening the door to immense suffering caused by those acting with malice, though they are not perfectly equivalent. The firearms industry injures through the deaths and injuries of the innocent, caused by a relatively small handful of people, while social media may subvert the moral and political life of a nation through the actions of many, sometimes in concert. But they may be equivalent in the misery and damage they cause.

The Section 230 safe harbor for information posted by third parties was meant to encourage the development and innovation of the then-nascent internet. The computational power and human effort needed to curate millions of internet posts, in accord with principles that apply to a newsroom, were beyond the capacity of computational systems at the time of Section 230's enactment. Holding internet social media companies to that liability meant drastically increasing their exposure to legal damages as to discourage their development to promote a wide exchange of information. This would have drastically reduced their ability to reach out and exchange information with the world. Although there were arguments that library or bookseller protections should apply to internet social media sites, at least one court held otherwise.[3] That ruling, from a New York trial court, placed every social media service on notice of potential and massive liability for data exchange and innovation in the online social media world. Because of this and other concerns, the U.S. Congress implemented the safe harbor provisions of Section 230 to protect

[1] 47 U.S.C. § 230 (2018).

[2] Protection of Lawful Commerce in Arms Act (PLCAA), Pub. L. No. 109–92 (2005).

[3] Stratton Oakmont, Inc. v. Prodigy Servs. Co., 1995 WL 323710 (N.Y. Sup. Ct. 1995).

and encourage evolution of these social media systems. These provisions supported the value platforms provide in making information widely available to everyone, regardless of their place in society or economic status.

Acting as the library and town square to the world, social media systems have been immensely successful. Discussions and the exchange of ideas cover the globe, embracing nearly every country and every political system on the planet. The issue is whether or not that development and innovation has come at the cost of injuries to the innocent from the malicious acts of others using social media to magnify the harm they do.

Hesiod describes—as vengeance for Prometheus and his technological gift of fire to mankind—Zeus's gift of Pandora's jar, which she opens, unleashing evil upon the world.[4] The rise of the internet parallels this parable as to the great benefits and risks of harm.[5] The internet topology, for all its benefits, effectively puts every user close to all the evils of which we could possibly conceive. Whether adult or child, old or young, temptation, corruption, and seduction become proximate in ways never before conceived nor anticipated. It can subvert traditional guardian roles that have protected people in the past. This may drive regulation even under the most solicitous regimes of freedom of expression.[6]

Then there is the prod of Artificial-Intelligence-driven recommender systems that introduce us to those evils we might otherwise ignore, possibly leading to corrupt conduct we might otherwise avoid. The debate over terrorists' use of these systems was momentarily settled for the U.S. Its Supreme Court decided, for the moment, social media systems have no responsibility in the case as presented. A greater showing of connection between the primary offenders and social media companies.[7] These decisions may influence regulation under general law relating to terroristic activities, depending on the engagement of the social media companies in the wrongful conduct, such as through its recommender and moderation systems. Or those companies may continue to operate with impunity under Section 230 for the conduct of others.

Like the offspring of the gift of fire—technology and civilization—the internet has a similar impact as a "vast library" of knowledge as well as a "sprawling mall" for goods and services; it was further characterized at the dawn

---

[4] HESIOD, THEOGONY; AND, WORKS AND DAYS (Oxford Univ. Press 1988) (700 BC).

[5] M. Mayer & J.E. Till, *The Internet: A Modern Pandora's Box?* 5 QUAL LIFE RES. 568, 568–71 (1996); Alex Trauth-Goik, *The Internet: Pandora's Box or The Horn of Cornucopia? The Digital Age Calls for an Internet Etiquette*, DIGIT. CULTURIST (Nov. 26, 2018), https://digitalculturist.com/the-internet-pandoras-box-or-the-horn-of-cornucopia-8f4dc53f6d48 accessed 3/8/2023 (on file with the *University of the Pacific Law Review*) ("For better or worse, the Internet has arrived as a force in our lives, and medical information and support constitute two of its more important uses.").

[6] Sian Cain, *Stephen Fry: Facebook and Other Platforms Should Be Classed as Publishers*, GUARDIAN (May 28, 2017), https://www.theguardian.com/technology/2017/may/28/stephen-fry-facebook-and-other-platforms-should-be-classed-as-publishers (on file with the *University of the Pacific Law Review*) (The comedian and early Internet adopter Stephen Fry, referencing Pandora's jar, noted this was a challenge that must be faced: "The dark side of the rise of machines and the sudden obsolescence of so many careers and jobs; the potential for crime, exploitation, extortion; suppression and surveillance; and even newer forms of cyberterrorism, give us the collywobbles and are challenges for certain. But we must understand that it is going to happen, collywobbles or not, because the lid is already off the jar. So the best we can do is keep the lid of the jar and let hope fly out.").

[7] Gonzalez v. Google LLC, 598 US 617, 622 (2023); Twitter v. Tannameh, 598 US 471, 505 (2023).

of the Internet in 1996 by the U.S. Supreme Court in *Reno v. American Civil Liberties Union*:

> From the publishers' point of view, it constitutes a vast platform from which to address and hear from a worldwide audience of millions of readers, viewers, researchers, and buyers. Any person or organization with a computer connected to the Internet can "publish" information. Publishers include government agencies, educational institutions, commercial entities, advocacy groups, and individuals. Publishers may either make their material available to the entire pool of Internet users, or confine access to a selected group, such as those willing to pay for the privilege. "No single organization controls any membership in the Web, nor is there any single centralized point from which individual Web sites or services can be blocked from the Web." [8]

In the 1996 challenge to accountability for content on the internet and commensurate efforts to regulate it, the U.S. Supreme Court said that "in the *absence of evidence to the contrary* [emphasis added], we presume that governmental regulation of the content of speech is more likely to interfere with the free exchange of ideas than to encourage it. The interest in encouraging freedom of expression in a democratic society outweighs any *theoretical but unproven* [emphasis added] benefit of censorship."[9]

That challenge was to the then-newly enacted CDA statute seeking to regulate access to sexually explicit materials by minors.[10] That same statute in Section 230 of the CDA also provided the safe harbor for social media sites for liability as to content posted on them by third party users. The legislature implemented it to shield social media seeking to try to control inappropriate content on their sites. Legislators felt that, absent such a shield, the nascent world of the internet, built on sharing information, would be throttled. Threats of litigation over internet content could not practically be monitored and controlled absent slowing internet development. Critics viewed early, conflicting court rulings and statutes regarding such liability as disincentivizing social media and internet expansion; no one wished to risk liability dependent on forum law that might be from anywhere in the world.[11]

Section 230 was immensely successful at insulating social media companies from liability for the content posted by others on their sites, no matter how injurious. Accountability for content well outside other speech protections,

---

[8] Reno v. Am. C.L. Union, 521 U.S. 844, 853 (1997).

[9] *Id.* at 885.

[10] Communications Decency Act, Pub. L. No. 104–104 (110 Stat.) 56 (1996).

[11] Stratton Oakmont, Inc. v. Prodigy Servs. Co., 1995 WL 323710 (N.Y. Sup. Ct. 1995) (supporting liability of the social media service); Cubby, Inc. v. CompuServe, Inc., 776 F. Supp. 135 (S.D.N.Y. 1991) (supporting immunity for the social media service); Smith v. California, 361 U.S. 147, 153 (1959) (held a bookseller cannot be held liable for "obscene" materials absent proof of knowledge such was obscene); 17 USC §108 of the US Copyright Act provides "the library exception" that protects libraries against copyright infringement claims absent knowledge of the infringement.

such as the First Amendment to the U.S. Constitution, rested only with the content originator, not the social media site, absent its involvement in the posts. Section 230 immunity removed the liability disincentive from expansion of the internet, especially for social media that seeks to expand information interaction among people. A string of unsuccessful plaintiff litigation over third party content and Section 230 demonstrated the strength and power of its shield.

In one of the first cases applying Section 230, a malicious, anonymous online prank via the AOL online service defamed someone and subjected him to public scorn. The defamed party received harassing telephone calls about claims that he sold "Naughty Oklahoma T-shirts" with offensive slogans about the 1995 Oklahoma City bombing of the Alfred P. Murrah Federal Building.[12] The U.S. Fourth Circuit Court of Appeals, in upholding the district court's dismissal of the case, noted:

> Congress made a policy choice ... not to deter harmful online speech through the separate route of imposing tort liability on companies that serve as intermediaries for other parties' potentially injurious messages ... The specter of tort liability in an area of such prolific speech would have an obvious chilling effect. It would be impossible for service providers to screen each of their millions of postings for possible problems. 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. Congress considered the weight of the speech interests implicated and chose to immunize service providers to avoid any such restrictive effect.[13]

A series of cases averring social media company liability for third party content followed, all dismissed under Section 230. Dismissal followed even where the social media system was alleged to have "culpably" assisted dissemination of the postings.[14]

Yet in *Reno v. ACLU,* the Supreme Court left open the possibility of reviewing the propriety of Section 230's shield, noting "in the *absence of evidence to the contrary*, we presume that governmental regulation of the content of speech

---

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

[13] *Id*..

[14] Universal Commc'ns Sys., Inc. v. Lycos, Inc., 478 F.3d 413, 422 (1st Cir. 2007) (although the Lycos system had 3rd party defamatory postings, it did not independently provide "culpable assistance" through registration and linking of bulletin boards as these features are "standard elements of web sites 'with [both] lawful and unlawful potential'… and hence, without more, cannot form the basis to find inducement."); Parker v. Google, Inc., 422 F. Supp. 2d 492, 501 (2006) *aff'd*, 242 Fed. Appx. 833 (3d Cir. 2007), *cert denied,* 522 U.S. 1156 (2008) (rejecting claims of defamation, invasion of privacy and negligence by Google for caching defamatory material on a third-party website and revealing it in Search; the court noted "It is clear that § 230 was intended to provide immunity for service providers like Google on exactly the claims Plaintiff raises here."); Perfect 10, Inc. v. CCBill LLC, 488 F.3d 1102, 1119 (9th Cir. 2007) (Section 230 provides immunity for social media systems for content posted by third-parties as to state law intellectual property claims; the court noted that the Digital Millennium Copyright Act provided similar immunity for claims of federal intellectual property infringement.).

is more likely to interfere with the free exchange of ideas than to encourage it. The interest in encouraging freedom of expression in a democratic society outweighs any *theoretical but unproven* [emphasis added] benefit of censorship." Developing evidence of interference with the free exchange and proven benefits from censorship may well shift the reasoning to permit a Section 230 amendment for social media companies.

The power of that shield was challenged in the aftermath of terrorist murders allegedly facilitated by social media. In *Gonzalez v. Google, Inc.* (2023)[15] and *Twitter v. Taamneh* (2023),[16] the U.S. Supreme Court addressed the possible liability of social media systems for deaths resulting from acts of terrorism. The social media systems allegedly aided and abetted the terrorists through use of these systems, including recommender services. The plaintiffs averred that the Antiterrorism and Effective Death Penalty Act of 1996, through its Section 2333,[17] created liability for content that was not shielded by Section 230. A unanimous court in *Taanmeh* did not address the Section 230 issue. It held that there was insufficient causation and the "allegations are insufficient to establish that these defendants aided and abetted ISIS [the terrorists] in carrying out the relevant attack" under Section 2333.[18] It found that the plaintiffs had not shown a "concrete nexus" between the social media systems and the terrorists. The court further remanded the *Gonzalez* case back to the appellate court to reconsider it in light of *Taamneh*.[19]

Though the Court did not address or settle any Section 230 immunity concerns, Justice Jackson's concurring opinion observed that the matter may still be open for review with greater factual development regarding the social media systems and their use. Such system review of operations may not invoke the common law principles applied in *Taanmeh*:

> Both cases came to this Court at the motion-to-dismiss stage, with no factual record. And the Court's view of the facts—including its characterizations of the social-media platforms and algorithms at issue—properly rests on the particular allegations in those complaints. Other cases presenting different allegations and different records may lead to different conclusions. The Court also draws on general principles of tort and criminal law to inform its understanding of §2333(d)(2). General principles are not, however, universal. The common-law propositions this Court identifies in interpreting §2333(d)(2) do not necessarily translate to other contexts.[20]

---

[15] Gonzalez v. Google, LLC, 598 U.S. 620, 622 (2023).

[16] Twitter, Inc. v. Taamneh, 598 U.S. 471, 505 (2023).

[17] Antiterrorism and Effective Death Penalty Act of 1996, 18 U.S.C.A. § 2333(a) ("Any national of the United States injured in his or her person, property, or business by reason of an act of international terrorism, or his or her estate, survivors, or heirs, may sue therefor in any appropriate district court of the United States and shall recover threefold the damages he or she sustains and the cost of the suit, including attorney's fees.").

[18] Twitter, Inc. v. Taamneh, 598 U.S. 471, 478 (2023).

[19] *Id*.

[20] *Id*. at 507.

This concurrence shows judicial rulings may yet play a role in deciding the scope of Section 230 immunity. Liability based on greater factual connections to causation would show that legislation may not be the sole route to changing U.S. regulation of social media.

Even though Section 230 immunity is limited to the United States, Facebook, X (formerly Twitter), TikTok, and other social media systems have been immensely successful and have expanded globally. Facebook, as of the third quarter of 2023, had 3.049 billion active users.[21] It is not alone; Table 1 lists the top six social media services and their active users, in billions:[22]

Facebook	2.96
YouTube	2.5
WhatsApp	2.0
Instagram	2.0
WeChat	1.3
TikTok	1.1

The immense scope and reach of these social media systems heightens concern over their impact. TikTok, first released in 2016, quickly grew in popularity to over a billion active users.[23] Given its scope, and as a service built and domiciled in the Peoples' Republic of China, it has generated immense controversy in the United States as a potential risk to national security, for reasons ranging from data sharing and spying to misinformation and recommender manipulation.[24]

Well before national security concerns arose, many members of the public objected to the impact of social media systems in a variety of areas. Section 230 immunity was contained within the Communications Decency Act of 1996 that obligated social media systems to monitor and block the exposure of children to "indecent" material online; the U.S. Supreme Court found such regulation was, based on the technology of the time, an onerous and improper burden on free expression, unlikely to be successful in protecting children. [25]

Section 230 immunity continues to be highly contentious for removing incentives for social media platforms to police their operations for inappropriate

---

[21] *Number of Monthly Active Facebook Users Worldwide as of 1st Quarter 2023*, STATISTICA, https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide// (last visited Jan. 21, 2024) (on file with the *University of the Pacific Law Review*) ("Table 1- top social media services and active monthly users- 2022 (in billions).").

[22] *Most Popular Social Networks Worldwide as of January 2023, Ranked by Number of Monthly Active Users*, STATISTICA, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (last visited July 16, 2023) (on file with the *University of the Pacific Law Review*).

[23] *Id*.

[24] Rachel Treisman, *The FBI Alleges TikTok Poses National Security Concerns*, NPR (Nov. 17, 2022), https://www.npr.org/2022/11/17/1137155540/fbi-tiktok-national-security-concerns-china (on file with the *University of the Pacific Law Review*); *Five Ways TikTok Is Seen As A Threat To National Security*, SEC. WK., https://www.securityweek.com/five-ways-tiktok-seen-threat-us-national-security/ (last visited July 16, 2023) (on file with the *University of the Pacific Law Review*); Zei Yang, *How China Takes Extreme Measures to Keep Teens Off Tik Tok*, MIT TECH. REV. (Mar. 8, 2023), https://www.technologyreview.com/2023/03/08/1069527/china-tiktok-douyin-teens-privacy/ (on file with the *University of the Pacific Law Review*).

[25] Reno v. Am. C.L. Union, 521 U.S. 844, 885 (1997).

conduct. Pressure for legislative and judicial limits will continue. Social media systems should prepare for potential changes that will impact their liability and how they program their systems.

II. WHAT TO DO?

Effective responses may be defined by the facts, law, and *ad hominem* bombast.[26] Regulation of social media and the internet have been impacted by all three, with the law moving towards greater restriction and the bombast getting louder. Political outrage comes from those who feel social media systems treat them unfairly and in a biased, negative manner, or promote misinformation, disinformation, and malicious information about them. In the United States, social media systems are whipsawed between competing interest groups regarding their content presentations and renditions. On one hand, some interest groups contend their views are suppressed by too many biased filtering systems coded against them. Other interest groups object to too few effective filtering systems to block inappropriate speech and information.

In *Reno v. ACLU* (1997), the Supreme Court cited the lower court's findings that no practical technology existed at that time to assure minors could not access indecent material online, short of a total ban on such ambiguous, ill-defined content. Yet by 2018, Congress amended the CDA's Section 230 Safe Harbor to limit immunity for social media sites when used to promote sex trafficking and child sexual exploitation, citing compelling state interests in public protection.[27] As such, third party content relating to sexual matters has been greatly restricted by social media platforms to avoid liability.

But while there appeared to be an immediate impact on such activity, the online marketplace may have recovered fairly quickly despite government claims of a "90%" reduction in such activity.[28] Some argue the act has had little impact overall in really reducing online sexual misconduct while endangering adult commercial sex workers by limiting their ability to vet clients.[29] This was attributed, in part, to the confusing and "vague, poorly defined carveouts to Section 230 [that] can spur platforms to over-moderate, with potentially disastrous effects for vulnerable people." Section 230 turns out to be "a sensitive dial": small adjustments can have "wide-reaching", and unanticipated, effects.[30] Analysts note that, in the immediate aftermath of the closure of the BackPage website—which offered sex services—that, rather than a long-term reduction in such sites,

---

[26] Anonymous ("If the facts are against you, argue the law. If the law is against you, argue the facts. If the law and the facts are against you, pound the table and yell like hell.").

[27] Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164, 132 Stat. 1253 (2018); Elizabeth M. Donovan, *Fight Online Sex Trafficking Act and Stop Enabling Sex Traffickers Act: A Shield for Jane Doe*, 52 CONN. L. REV. 85, 85–122 (2020).

[28] Glenn Kessler, *Has the Sex-Trafficking Law Eliminated 90 Percent of Sex-Trafficking Ads?*, WASH. POST (Aug. 20, 2018), https://www.washingtonpost.com/politics/2018/08/20/has-sex-trafficking-law-eliminated-percent-sex-trafficking-ads/ (on file with the *University of the Pacific Law Review*).

[29] Danielle Keats Citron & Quinta Jurecic, *FOSTA's Mess*, 26 VA. J. L. & TECH. 1 (2022–2023).

[30] Quinta Jurecic, *The Politics of Section 230 Reform: Learning from FOSTA's Mistakes*, BROOKINGS 10 (Mar. 1, 2022), https://www.brookings.edu/articles/the-politics-of-section-230-reform-learning-from-fostas-mistakes/ (on file with the *University of the Pacific Law Review*).

displacement opened new sites, outside U.S. jurisdiction, to meet market demand; they further suggest a similar shift of safety and market power from sex workers themselves to customers, possibly increasing risk while not reducing activity.[31]

Yet others reject "the claim by sex workers and sex worker rights advocates that the alleged burdens FOSTA-SESTA puts on those who self-report as freely choosing to work in the sex trade outweigh the potential benefit—fewer sex-trafficked people."[32] The FOSTA outcomes serve as an ongoing case study into what may or may not be an effective effort to curb social media misconduct and what may lead to worse outcomes.

Similar risks may accompany further efforts to amend Section 230 immunity to regulate social media while having no real impact on online misconduct. This again produces the problem faced by the Supreme Court in *Reno v. ACLU*, where a compelling, laudable purpose still cannot find an effective solution to justify limiting rights of free expression. One example of damage from censorship laws is the efforts of the Florida legislature[33] to remove offensive books from school libraries; this led to the removal of a children's book by Roald Dahl along with over 170 others.[34]

The law alone will have difficulty providing a just and effective solution. Use of legal liability disincentives may lead to unintended consequences that outweigh any actual benefit.

A. The Possibility of Social/Fact/Technical Solutions to Risks of Harm

Online misconduct leads to grave injuries; the challenge is devising solutions that maintain the benefits of social media while reducing those injuries. Primary regulatory impulses, like those in FOSTA-SESTA, are to attack social media systems' deep pockets and hold them accountable for others' misconduct, rather than attacking those individuals directly engaged in the misconduct. This

---

[31] Mike Tobias**,** *How the Backpage Shutdown Impacted the Commercial Sex Industry and Trafficking*, NEB. PUB. MEDIA (Aug. 21, 2018), https://nebraskapublicmedia.org/en/news/news-articles/how-the-backpage-shutdown-impacted-the-commercial-sex-industry-and-trafficking/ (on file with the *University of the Pacific Law Review*); Helen Shuxuan Zeng, Brett Danaher & Michael D. Smith, *Internet Governance Through Site Shutdowns: The Impact of Shutting Down Two Major Commercial Sex Advertising Sites*, 68(11) MGMT. SCI. 8234 (2022).

[31] *See* FRANK MORGAN, *Displacement of Crime*, THE ENCYCLOPEDIA OF THEORETIC AL CRIMINOLOGY (Mar. 2014) (This accords with social displacement theory that posits the unintended effect of crime prevention may simply move criminal activity to other places and times, thus failing to reduce crime overall. Prevention efforts must take this into account as to accomplish an actual reduction in criminal activity.); Elizabeth M. Donovan, *Fight Online Sex Trafficking Act and Stop Enabling Sex Traffickers Act: A Shield for Jane Doe*, 52 CONN. L. REV. 85, 85–122 (2020).

[32] Elizabeth M. Donovan, *Fight Online Sex Trafficking Act and Stop Enabling Sex Traffickers Act: A Shield for Jane Doe*, 52 CONN. L. REV. 85, 85–122 (2020).

[33] FLA. STAT. § 1006.40 (2021).

[34] Lisa Tolin, *These 176 Books Were Banned in Duval County Florida*, PEN AM. (Dec. 6, 2022), https://pen.org/banned-books-florida/ (on file with the *University of the Pacific Law Review*); Katy Waldman, *What Are We Protecting Children From By Banning Books,* NEW YORKER, Mar. 10, 2023; *Florida County Ranks High on List of School Book Bans*, FIRST COAST NEWS (Sept. 22, 2022), https://www.firstcoastnews.com/article/news/education/duval-county-ranks-high-on-list-of-school-book-bans-district-claims-otherwise/77-d313e392-4969-4157-9d55-ce4c9a224610 (on file with *University of the Pacific Law Review*). *See* Sally Percy, *Six Leadership Lessons from Roald Dahl Books*, FORBES MAGAZINE (Sept. 13, 2019) (detailing the virtues shown in Dahl's books).

may not produce the results desired and lead to other injuries while damaging access to information some may deem vital.

U.S. Justice Louis Brandeis' Counter-Speech doctrine—which posits that the proper solution to bad speech is more speech— should be considered here.[35] But while the democratic power of social media systems supports the counter-speech doctrine—as anyone can now become a mass-publisher for relatively little cost—it does not fully account for the maturity, skill, familiarity, and access to resources needed to adequately respond to malice. Nor can it account for the damage done by misinformation across many groups of different levels of education, discretion, maturity, and access to counter-vailing discussion. These inequalities indicate a need for some means of addressing malice and ignorance on social media. For counter-speech to work with these areas, there must be a robust community willing to counter malicious speech. The challenge is how to accomplish this. A community needs a culture of engagement willing to accept the risks of countering others' views. This includes such things as hateful online attacks, "doxing"—by which personal files are sent out into the world—and other malicious activities possible in the online world. The relative anonymity that currently exists makes the risk of such malicious conduct seemingly less, and the bad conduct all the more likely. Those engaging in counter-speech online would need to be prepared for such attacks.

Criminological models to reduce criminogenic activity may offer some guidance as to participatory counter-speech and online conduct generally. Social Organization Theory indicates that it is the absence or breakdown of community and civic relationships that leads to an increase in criminal activity; the challenge is identifying what relationships create the greatest risks.[36] For online systems, the risks abound given the potential weak connections between online users. Opportunity Theory, Routine Activity Theory, and Situational Crime Prevention Theory, among others, indicate general frameworks for protecting people from information injury. They can provide this while securing the free exchange of ideas and benefits of the internet.[37] The benefits of the internet machine include social media and its ability to easily connect people to one another across broad and diverse interests and demographics.

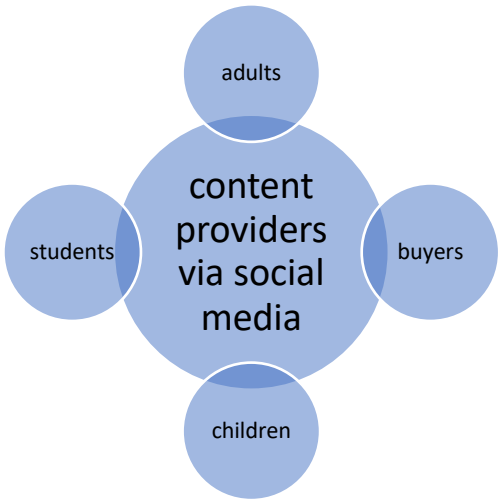
---

[35] Whitney v. California, 274 U.S. 357 (1927); Robert D. Richards & Clay Calvert, *Counterspeech 2000: A New Look at the Old Remedy for "Bad" Speech*, 2000 BYU L. REV. 553 (2000); Robert D. Richards & Clay Calvert, *Counterspeech 2000: A New Look at the Old Remedy for "Bad" Speech*, 2000 BYU L. REV. 553 (2000).

[36] CHARIS KUBRIN & JAMES WO, *Social Disorganization Theory's Greatest Challenge: Linking Structural Characteristics to Crime in Socially Disorganized Communities*, *in* THE HANDBOOK OF CRIMINOLOGICAL THEORY 121–36 (Wiley Publ'g 2015).

[37] *Why Crimes Occur in Hot Spots*, NAT'L INST. OF JUST. (Oct. 13, 2009), https://nij.ojp.gov/topics/articles/why-crimes-occur-hot-spots (on file with the *University of the Pacific Law Review*).

Consider those involved in online activity via a social media intermediary system, whether Google, Facebook, or Twitter:



The immediacy of social media in connecting Party A to Party B brings the benefits of efficiency and the detriments of unmediated contact. Under Routine Activity Theory, a motivated offender attacks a suitable target where there is no effective guardian for the target.[38] Situational Crime Prevention Theory posits crime is deterred by reducing opportunities. Crime Opportunity Theory similarly posits that a reduction in opportunities reduces crime. All of these indicate a possible role for expanding protections at the central mediating hub of a social media system. This can protect social media users, regardless of their social media knowledge or sophistication.

Consider some potential solutions, including supporting legislation, that may lead to accountability for misconduct or misinformation:

1) Amend Safe Harbor immunities to encourage effective technical solutions that curate content and reduce harmful activity. These may include Artificial Intelligence and machine learning algorithmic filtering. This may require new and different types of immunity for content-curating technology, which may still produce erroneous information.
2) Provide mechanisms that reduce the anonymity of those perpetrating online misconduct so they, as the primary wrongdoers, are held accountable as deterrents to online misconduct; this has the potential to conflict with long-held principles supporting anonymous speech, as a means of presenting unpopular ideas to others.

[38] Lawrence E. Cohen & Marcus Felson, *Social Change and Crime Rate Trends: A Routine Activity Approach*, 44 AM. SOC. REV. 588, 588–605 (Aug. 1979).

3) Provide effective personal autonomy choices where individuals have access to technical tools that let *them* choose content and protect themselves from online deviance; this would include both technical tools and instruction on how to effectively use them.
4) Provide support for algorithmic systems that connect people in the online world in positive, evaluative ways, which can mediate and bridge disagreements and differing views between people, rather than promoting or attacking them.

Such melding of technical fact reform and legal duty may be much more effective in reducing detriments while maintaining the benefits of social media systems. The implementation of such efforts may be difficult, as the devil is always in the details.

B. Possible Implementations of Internet Social Media Protections

1. Amend Safe Harbor Immunities to Encourage Effective Technical Solutions, Such as Artificial Intelligence and Machine Learning Algorithmic Filtering, to Curate Content and Reduce Harmful Activity From "Bad" Speech and the Blocking Of Legitimate Speech.

Modified conditions for immunity might provide protection for a social media system that implements algorithmic vetting and curating of content. Such a system analyzes the content against a "fair" ruleset, rather than simply banning content without analysis. There are concerns that liability will drive social media systems to ban far more than prohibited activity as blanket blocking is easier and legally safer than more discreet approaches. This may, in turn, result in reducing the scope of legitimate, democratic discourse on a variety of important issues.

AI systems provide greater analytical nuance, but are not completely reliable, given the ongoing evolution of their own algorithms and the nature of speech and discussion—especially where "spoofing" of media is used to bypass curation systems. Regulations must accommodate false positive and false negative rates in the curation process, as these systems are never perfect. They should provide safe harbor protections for systems that seek to balance barring harmful, illegal content with permitting dissemination of permissible content.

Modified "good faith" safe harbor provisions would need to:

i) Define the scope of "good faith" curation systems,
ii) Set out rules to qualify curation systems for the legislative safe harbor, or
iii) Provide a dynamic mechanism by which changes in algorithmic power, misinformation, and spoofing capabilities can be monitored and incorporated into dynamic curation systems.

Such a regulatory system would minimize damage to permissible speech while providing some controls for misinformation. Hard-coding, filtering algorithms, supervised machine learning, and unsupervised machine learning all

have advantages and disadvantages in balancing the cost of the filtering system against its effectiveness. Protection for false positives and false negatives in filtering will be necessary, as the systems cannot be perfect. Through ongoing review, the system will need to be able to respond to changes in systems and information practices to remain effective in limiting bad speech and damage to free speech.

Hybrid systems involving regular human review of AI output may play a role in this process. But such vetting may be prohibitively expensive. And the engagement of human reviewers has risks to those reviewers as they are exposed to the vile content, such as sexual abuse, hatred, and violence, on a regular basis.[39]

2. Provide Mechanisms That Reduce the Anonymity of Those Perpetrating Online Misconduct so They, As the Primary Wrongdoers, Are Held Accountable ss Deterrents to Online Misconduct.

Accountability for the primary miscreants in social media could reduce online misconduct. The opportunities for anonymous speech can lead to online misconduct with impunity and disinhibition. Citron has suggested reducing anonymity online as to hold primary perpetrators accountable.[40] Davenport suggests that anonymity online "puts the fabric of our society at risk."[41]

Yet anonymous speech has a long history in the United States and is protected under the First Amendment. Anonymous communication and associational[42] aspects of privacy and free speech protect against disclosure of authorship of anonymous speech. "Anonymity is a shield from the tyranny of the majority."[43] The Supreme Court reasoned, *"*[i]t thus exemplifies the purpose behind the Bill of Rights, and of the First Amendment in particular: to protect unpopular individuals from, retaliation—and their ideas from suppression—at the hand of an intolerant society.*"* Bodle notes the protection of anonymity extends globally, with the tension between the needs of some for anonymity and the desires of private social media companies for real identification for commercial benefit.[44] The Association for Progressive Communication asserts that "[a]nonymity is fundamental for the full exercise of the right to freedom of expression, as enshrined

---

[39] *See* Karen Hao & Deepa Seetharaman, *Cleaning Up ChatGPT Takes Heavy Toll on Human Workers*, WALL STREET J. (July 24, 2023), https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483 (on file with the *University of the Pacific Law Review*) (explaining that the use of human reviewers in the testing and tuning of ChatGPT and other systems itself became controversial due to the impact on them, which has in turn led to a petition to the Kenyan Legislature for relief and legal action against a social media company for the toll on the workers).

[40] Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61 (2009) (arguing that Section 230's immunity for under-filtering illegality should be conditioned on a "duty of care" negligence standard, which for instance would include the facilitation of traceable anonymity so perpetrators could be caught and sued).

[41] David Davenport, *Anonymity on the Internet: Why the Price May Be Too High*, 45(4) COMMC'NS ASS'N FOR COMPUTING MACH. 33–35 (2002).

[42] NAACP v. Alabama *ex rel.* Patterson, 357 U.S. 449, 462 (1958).

[43] McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 334 (1995).

[44] Robert Bodle, *The Ethics of Online Anonymity or Zuckerberg Vs. "Moot"*, 43 COMPUTS. & SOC'Y 1, 22–25 (May 2013).

in Article 19 of the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights."[45]

Many regimes ban anonymity online—from unsigned online text posts, to a requirement that image and video files contain identifying watermarks of provenance. Anonymity protects those posting from punishment for criticism of those regimes just as it hides miscreants from accountability. Yet compelling state and community interests may overcome the need for protections of online speech, This can be soeven under the strictest of scrutiny, if the damage to vital interests becomes too great without reasonable means of correction.

There may be other possibilities inherent in the democratic and increasingly egalitarian nature of social media and internet technologies. Venkataraman suggests enabling individual power over data technologies can give us the protection needed and choice desired.[46] But she is concerned a lack of imagination may hobble these possibilities. She notes the analysis of Jonathan Zittrain that, "We basically have two problems in digital governance: one, not knowing what we want and, two, not trusting people to give it to us."[47] Offering options may be one means of focusing on effective solutions, and suggests technology permit greater individual control over information systems and encourage people to bridge their differences.

3. Provide Effective Personal Autonomy Choices Where Individuals Have Access to Technical Tools To Choose Content Access and Protect Themselves From Online Deviance.

Personal autonomy is an inherent part of the dignity of each person that is imperative under the foundations of ethics set out by Immanuel Kant.[48] Rather than requiring social media systems to filter and curate content, social media systems and third parties may offer personal curation systems by which users can choose and direct which content they are given. Filtering systems typically block content, but a mediated algorithmic system may seek out relevant and useful information. System presets may be designed to make customization easy for everyone and make transparent the ways in which content is curated. This might include inclusion and exclusion lists a user can peruse to modify the curation process. This is similar to spam filtering, which lets a user check and change what it has blocked. An equivalent system of blacklisting blocked material or whitelisting permitted material may be another option. This is similar to that used for content protection software to limit children's access to some materials.

---

[45] *The Right to Freedom of Expression and the Use of Encryption and Anonymity in Digital Communications,* ASS'N FOR PROGRESSIVE COMMC'NS 1 (Feb. 2015), https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Communications/AssociationForProgressiveCommunication.pdf (on file with the *University of the Pacific Law Review*) ("The right to freedom of expression and the use of encryption and anonymity in digital communications.").

[46] Bina Venkataraman*, Opinion: A Better Kind of Social Media Is Possible — If We Want It*, WASH. POST (Mar. 6, 2023), https://www.washingtonpost.com/opinions/2023/03/06/social-media-future-regulation-imagination/ (on file with the *University of the Pacific Law Review*).

[47] *Id*.

[48] IMMANUEL KANT ET AL., GROUNDWORK FOR THE METAPHYSICS OF MORALS § 4:429 (Oxford Univ. Press 2019) (1785).

4. Provide Support for Algorithmic Systems That Connect in Positive, Evaluative Ways, Bridging Disagreements Between People.

Social media systems can be incentivized to offer system options that let people connect and discuss matters between them in less hate-filled ways. Ovadya and Thorburn suggest that the disunity online can be overcome by systems that allow online "bridging systems." Online connections between people may be achieved through algorithms and artificial intelligence systems directed to that end.[49]

Such systems would include a focus on recommender systems on social media and their algorithmic foundations. They would be paralleled by introducing human-in-loop systems of software for conducting civic forums and human-facilitated group deliberation. Such systems could also be integrated into existing social media algorithms, through which a user could choose a path of engagement with others. This is an alternative to algorithmic recommendations that engage the user with self-reinforcing interests, beliefs, and preferences.

This alternative also offers the opportunity for counter-speech to play its role in the process. It reflects the utility of the counter-speech doctrine. This may be accompanied by reform of legal duties that may impact and improve social media through regulation.[50] Reform of such legal obligations may include modification of the safe harbor provisions currently protecting social media. This may build an incentive for social media systems or cause them to limit their services due to technological limitations and fear of liability. [51] This can apply to injuries to children from online misinformation and predation, for which damages may be immense. And the interest in protecting children is a compelling one, and one of sympathy to any judge or jury.

Citron suggests this may be done by creating duties of care for social media and limiting safe harbor immunities where knowing and intentional support of bad online conduct is shown.[52] This requires very careful legislative drafting to provide practical protection while not destroying the benefits social media online offer. The facts of these technologies, though, are changing and they may offer new solutions that protect the best of what these information technologies offer.

---

[49] Aviv Ovadya & Luke Thorburn, *Bridging Systems: Open Problems for Countering Destructive Divisiveness Across Ranking, Recommenders, and Governance*, (Cornell Univ. Working Paper, Paper No. 2301.09976), https://doi.org/10.48550/arXiv.2301.09976 (on file with the *University of the Pacific Law Review*).

[50] Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CAL. L. REV. 1805 (2010); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 24 (Harv. Univ. Press 2014) (proposing a Section 230 carveout for sites that principally host nonconsensual pornography and cyber stalking).

[51] Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 406 (2017) (offering statutory language to operationalize a duty of care in a "reasonable steps" approach); Danielle Keats Citron & Mary Anne Franks, *The Internet as Speech Machine and Other Myths Confounding Section 230 Reform*, 2020 U. CHI. LEGAL F. 45 (2020).

[52] Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. REV. 713 (2023).

C. Things to Come

The San Mateo County Board of Education and school officials have filed suit in U.S. District Court alleging YouTube, TikTok and Snapchat have, through intentional design to lure and keep viewers, are damaging the mental health of children.[53] The plaintiffs assert liability under public nuisance, negligence, racketeering and corrupt influences, conspiracy, gross negligence, and unfair competition for injuries to the mental health of children. With passion, they detail their cause in a 107-paged complaint:

> 2. The full scale and depravity of the YouTube, TikTok, and Snap companies' conduct in this case may only be fully understood after this Court issues injunctive relief addressing what President Biden recently referred to, in his 2023 State of the Union address, as the "experiment" that Defendants and other major social media companies "are running on our children for profit." In addition to injunctive relief restraining Defendants and their co-conspirators from further engaging in their unlawful conduct, this action seeks to recover San Mateo County schools' costs to address the youth mental health crisis caused by the YouTube, TikTok, and Snap companies' conduct, and to compel Defendants to disgorge the profits of their unlawful conduct.[54]

The alleged manipulation is done, claim the plaintiffs, through the use of "the most advanced" Artificial Intelligence Technology in the world, which make their sites "addictive and to deliver harmful content to youth."[55]

This is a different approach to liability that may or may not be barred by Section 230. It may circumvent Section 230 because, although tendering user-posted content, the social media systems themselves are deciding what to make available to the users, including children. That agency may, perhaps, hold social media systems accountable based on the recommender systems they use.

Almost 200 other school districts and parents themselves have joined in this litigation against the social media giants asserting similar and related liability claims for the impact of those social media platforms on their children.[56] They assert that responding to and training for disciplinary issues, such as cyberbullying, takes their time away from the teaching of their students and they should be compensated. The defendant social media platforms, asserting they operate with concern for children, moved to dismiss the complaint, arguing that Section 230 still applies given the wrongs originate in the content of third parties. The parents' claims assert defective product design and negligence; they assert these are not

---

[53] San Mateo Cnty. Bd. of Educ. et al. v. YouTube, LLC. et al., No 3:23-cv-01108, at *1–107 (N.D. Cal. Mar. 13, 2023).

[54] *Id*.

[55] *Id*.

[56] Sara Randazzo & Ryan Tracy, *Schools Sue Social-Media Platforms Over Alleged Harms to Students*, WALL STREET J. (July 23, 2023), https://www.wsj.com/articles/schools-sue-social-media-platforms-over-alleged-harms-to-students-ebca91a5 (on file with the *University of the Pacific Law Review*).

barred by Section 230. This may parallel the reasoning in *Lemmon v. Snap* that Snap was not protected by Section 230 where two boys were killed while using a Snap SpeedFilter app feature that recorded their speed while using the service.[57] The court said:

> Snap, Inc. was sued for the predictable consequences of designing Snapchat in such a way that it allegedly encouraged dangerous behavior. Accordingly, the panel concluded that Snap, Inc. did not enjoy immunity from this suit under § 230(c)(1) of the CDA.[58]

The State of New Mexico filed its suit against Meta in December 2023 alleging damage for Meta's enabling and failure to prevent sexual exploitation materials from affecting New Mexico residents.[59] This was allegedly promoted by its systems algorithms, though Meta was aware of the sexual exploitation.. Documents referenced in the litigation indicate what was called a "historical reluctance" regarding the protections in this space. [60] Knowledge of injuries' cause, and failure to mitigate, may open the door to both punitive damages and punitive corrective legislation.

We should not wait for judicial incentivization for the protection of the innocent and opposition to the malicious. The social media companies' best interests are to collaborate on legislative remedies that create accountability for injuries through use of their systems. This offers protection when they act responsibly and effectively to protect those they may injure while preserving the exceptional benefits of their systems.

III. CONCLUSION

Not everything was lost from Pandora's jar. She, and humankind, were able to keep Hope. Hope came to support all humans in the face of all the evils unleashed by the malice of the gods. The internet and social media have done much good in the world, just as some have used it to unleash evil. There is hope that effective solutions will protect the good while reducing the bad. These are not solely legal solutions, but legal-technical ones that can best do this. We must support the innovation to develop these to preserve the great good offered. Hope for the best is something we must all have, though we must be very careful as we go forward surely and vigorously.

---

[57] Lemmon v. Snap, Inc., 995 F.3d 1085, 1087 (9th Cir. 2021).

[58] *Id*.

[59] State of New Mexico *ex rel*. Paul Torrez, Att'y Gen. v. Meta Platforms, et al, No. D-101-CV-2023-02838, (D. N.M. filed Dec. 18, 2023).

[60] Barbara Ortutay, *Court Documents Underscore Meta's 'Historical Reluctance' to Protect Children on Instagram: Newly Unredacted Documents from New Mexico's Lawsuit Against Meta Underscore the Company's Historical Reluctance to Keep Children Safe on its Platforms*, ASSOCIATED PRESS, Jan. 17, 2024.

* * *