The Business, Entrepreneurship & Tax Law Review

Volume 2
Issue 2 *Symposium: Innovation in Media and Entertainment Law*

Article 4

2018

Beyond Section 230: Liability, Free Speech, and Ethics on Global Social Networks

Brett G. Johnson

Beyond Section 230: Liability, Free Speech, and Ethics on Global Social Networks

Brett G. Johnson, PhD[*]

ABSTRACT

This article conducts a comparative analysis of intermediary liability laws regarding harmful speech in nine liberal-democratic polities. Harmful speech here is defined in terms of the ability of user-generated content ("UGC") to lead to individual physical harm (e.g., threats, incitement to violence), individual relational harm (e.g., defamation), or individual reactive harm (e.g., hate speech), as well as its potential to lead to social harm (e.g., fake news). The purpose of this comparative analysis is to distill a set of common principles upon which the concept of "platform ethics"—ethical duties that digital intermediaries owe to their users and to society—can be based. Conceptualizing platform ethics is incredibly important today as major social networks remain indispensable tools for democracy despite waning public trust stemming from recent major scandals.

[*] Assistant Professor of Journalism Studies at the University of Missouri School of Journalism. Professor Johnson holds a MA in Journalism from the University of Iowa and a PhD in Mass Communication from the University of Minnesota. Johnson's research involves both traditional legal research and the sociological study of law. Johnson would like to thank Dr. Amy Kristin Sanders for her guidance with portions of this project. Some portions of this article's introduction and background discussions are taken, in whole or in part, from the author's dissertation, Brett G. Johnson, *The Free Speech Balancing Act of Digital Intermediaries: An Explication of the Concept of Content Governance* (2015) (unpublished PhD dissertation, University of Minnesota) (on file with author).

I. INTRODUCTION

In 2016, Americans realized that social media can be weaponized to spread disinformation among highly targeted groups of users and sow chaos and confusion of varying types and degrees.[1] The locus of the fracas is on social media platforms, defined by media scholar Tarleton Gillespie as "sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices."[2] Trolls and provocateurs continue to plague these platforms with speech that ranges from merely vile to outright abuse and harassment.[3] Nevertheless, social media platforms remain a powerful tool for individuals to participate in a global public discourse and create change within their communities.[4] So, does that mean that society must take the bad with the good when it comes to social media? Or do social media platforms owe us a bit more?

In a 2017 article, I posed the question of whether social media platforms have a moral duty to prevent harm caused by the speech of others, promote freedom of expression, or some ideal combination of the two.[5] Although I framed the first possible moral duty in that article in terms of harms directed against specific individuals, it is possible for this question to be applied more broadly to other types of harmful content common in our world today: fake news, hate speech directed at a group of people, or speech glorifying terrorism. The issue of whether intermediaries might have a moral duty to promote freedom of expression is a tougher issue. Social networking platforms such as Facebook, Twitter, and YouTube (among others) sell users on the promise of freedom of expression and the allure of fame that might come from a user's content becoming widely popular (i.e., "going viral"), with the express purpose of commodifying and profiting off of users' content.[6] Accordingly, I have argued that these digital intermediaries have developed an "aggregational theory of freedom of expression"; primacy is placed on the capacity, magnitude, or

1. *See* Eric Westervelt, *How Russia Weaponized Social Media with 'Social Bots'*, NPR (Nov. 5, 2017, 8:06 AM), https://www.npr.org/2017/11/05/562058208/how-russia-weaponized-social-media-with-social-bots.

2. Tarleton Gillespie, *Governance of and by Platforms*, *in* SAGE HANDBOOK OF SOCIAL MEDIA 254 (Jean Burgess, Thomas Poell & Alice E. Marwick eds., 2017).

3. *See* Lee Rainie, Janna Anderson & Jonathan Albright, *The Future of Free Speech, Trolls, Anonymity and Fake News Online*, PEW RES. CTR (Mar. 29, 2017), http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

4. *See, e.g.*, Taso G. Lagos, Ted M. Coopman & Jonathan Tomhave, "*Parallel Poleis": Towards a Theoretical Framework of the Modern Public Sphere, Civic Engagement and the Structural Advantages of the Internet to Foster and Maintain Parallel Socio-Political Institutions*, 16 NEW MEDIA & SOC'Y 398 (2014); Andrew J. Flanagin, Craig Flanagin & Jon Flanagin, *Technical Code and the Social Construction of the Internet*, 12 NEW MEDIA & SOC'Y 179 (2010).

5. Brett G. Johnson, *Speech, Harm, and the Duties of Digital Intermediaries: Conceptualizing Platform Ethics*, 32 J. MEDIA ETHICS 16 (2017) [hereinafter *Speech, Harm and the Duties of Digital Intermediaries*].

6. *See, e.g.*, Louis Leung, *User-Generated Content on the Internet: An Examination of Gratifications, Civic Engagement and Psychological Empowerment*, 11 NEW MEDIA & SOC'Y 1327 (2009); Ute Schaedel & Michel Clement, *Managing the Online Crowd: Motivations for Engagement in User-Generated Content*, 7 J. MEDIA BUS. STUD. 17 (2010). *Cf.* Scott Wright, *Politics as Usual? Revolution, Normalization and a New Agenda for Online Deliberation*, 14 NEW MEDIA & SOC'Y 244 (2012).

potential of users to speak on platforms, and the quality or importance of that speech is overlooked.[7]

The fact that social media platforms are global operations complicates the ethical debate surrounding the promotion of free expression. In the United States', social networks have a First Amendment right to manage users' content as they please, and users are unable to make any constitutional claims against social networks for removing—some might say "censoring"—their speech.[8] However, this does not mean that these intermediaries should shrug off any moral duty toward promoting and practicing the values of freedom of expression among its users. Certainly, scholars have argued that the promotion of freedom of expression is not inherently a moral duty, trending instead toward hedonism.[9] However, by promoting and practicing the values of freedom of expression, digital intermediaries could act as a model for society to tolerate and critically engage with extreme and challenging ideas.[10] Furthermore, since these intermediaries are borne out of the exceptional American ethos for freedom of expression, one could argue they have a moral obligation to promote an American vision of freedom of expression in their global operations, particularly in countries that repress this basic human right.[11] However, one could just as easily argue that the privilege to operate globally requires social media companies to honor the social and cultural (to say nothing of legal) norms of the countries in which they do business.[12] Therefore, understanding how the harms of speech are defined, categorized, weighted, and punished in various parts of the world is crucial for social media platforms to operate ethically across the globe.

This article is as much about ethics as it is about law—as much about self-regulation as it is about government-imposed regulation. Indeed, it is about the connection between these poles; laws often can reflect the moral norms of the people that pass them.[13] I propose the best way to distill any moral principles that can be applied to how social media platforms *should* govern harmful content is to examine the laws that *mandate* how social media platforms govern harmful content. I argue

---

7. Brett G. Johnson, *Facebook's Free Speech Balancing Act: Corporate Social Responsibility and Norms of Online Discourse*, 5 J. MEDIA L. & ETHICS 17 (2016) [hereinafter *Facebook's Free Speech Balancing Act*].

8. *See, e.g.*, Cyber Promotions, Inc. v. AOL, Inc., 948 F. Supp. 436, 456 (E.D. Pa. 1996) (holding that AOL's email service was not the "functional equivalent" to a public forum. In other words, AOL was not acting as an agent supplying a forum for communication that state actors would normally make available. The court also held that AOL, unlike cable systems, did not control the "critical pathway" of communication, and thus an individual did not have a valid claim under 42 U.S.C. § 1983 that AOL was operating as a state actor in censoring his speech).

9. *See* Don E. Tomlinson, *Where Morality and Law Diverge: Ethical Alternatives in the* Soldier of Fortune *Cases*, 6 J. MASS MEDIA ETHICS 69 (1991).

10. *See* Brett G. Johnson, *Networked Communication and the Reprise of Tolerance Theory: Civic Education for Extreme Speech and Private Governance Online*, 50 FIRST AM. STUDIES 14 (2016).

11. Ethan Zuckerman, *Intermediary Censorship*, *in* ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE 71, 82–83 (Ronald Deibert et al. eds., 2010).

12. *See, e.g.*, Patrick L. Plaisance, *The Mass Media as Discursive Network: Building on the Implications of Libertarian and Communitarian Claims for News Media Ethics Theory*, 15 COMM. THEORY 292 (2005) (arguing that libertarianism has little, if any, moral justification as an ethical framework); Michael Perkins, *International Law and the Search for Universal Principles in Journalism Ethics*, 17 J. MASS MEDIA ETHICS 193, 205 (2002) (arguing that "the Western orientation of human rights treaties and the concept of free expression they contain can become problematic.").

13. Steven Shavell, *Law Versus Morality as Regulators of Conduct*, 4 AM. L. & ECON. REV. 227, 247 (2002).

that such an approach can reveal how various polities conceive of the ideal relationship between platforms and their users. Therefore, this article conducts a comparative analysis of intermediary liability laws from multiple countries.[14] In particular, this study looks at liability in the context of user-generated content ("UGC") that has the potential to cause harm to individuals, groups, or society at large.[15]

Part II defines in greater detail what is meant by harms to individuals, groups, or society at large caused by UGC.[16] However, because these harms are purposefully broad, some exclusions must be made lest the focus of this analysis become muddied from the outset.

First, laws dealing with vicarious intermediary liability in copyright infringement are excluded. Although copyright infringement is considered a moral harm against rights-holders in some countries,[17] it is primarily recognized as a commercial harm.[18] Thus, intermediary liability *vis-à-vis* copyright infringement is not an area of law from which moral duties to prevent harms to individuals, groups, and society at large can be distilled through comparative analysis.

Second, the distribution of images of child abuse (what is colloquially referred to as "child pornography") is excluded from the analysis. This type of content consists of contraband that is categorically classified as criminal across the globe, and a common feature of intermediary liability laws is that intermediaries are legally obligated to stanch the distribution of this material if they become aware of it.[19] Thus, this area of law leaves virtually no ability to compare and contrast nuanced legal, and therefore ethical, principles.

Third, laws regulating the management of personal data are excluded. This area of law is incredibly vast and rapidly evolving, and numerous scholarly articles have focused on comparative studies of data protection laws, especially those pertaining to the notion of a "right to be forgotten."[20] Including this area of law within this comparative analysis would add little to our understanding of the role of digital intermediaries in data protection. Furthermore, a conceptual difference exists between these two harms: the type of harm resulting from mismanagement of personal

14. A supranational polity (the European Union) is included among the countries examined.

15. User-generated content ("UGC") is defined herein as any message or media product created informally by individuals or groups vis-à-vis an online platform that facilitates such creation. *See, e.g.*, Ramon Lobato, Julian Thomas & Dan Hunter, *Histories of User-Generated Content: Between Formal and Informal Media Economies*, 5 INT'L J. COMM. 899, 900 (2011).

16. *See infra* notes 28–88 and accompanying text.

17. *See, e.g.*, Berne Convention for the Protection of Literary and Artistic Work art. 6*bis* (1), Sept. 28, 1979, http://www.wipo.int/wipolex/en/treaties/text.jsp?file_id=283693 ("Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation.").

18. RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 49 (1992).

19. *See, e.g.*, 47 U.S.C. § 230(e)(1) (2018) (providing that digital intermediaries will not have immunity from liability for third-party content that violates federal laws prohibiting the distribution of images of sexual exploitation of children).

20. Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, 2014 E.C.R. I-317, http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&doclang=EN (interpreting Directive 95/46/EC to hold that the "'right to be forgotten' . . . override[s] the legitimate interests of the operator of the search engine and the general interest in freedom of information."). For more context on this issue, *see, e.g.*, Jane E. Kirtley, *"Misguided in Principle and Unworkable in Practice": Time to Discard the* Reporters Committee *Doctrine of Practical Obscurity (and Its Evil Twin, the Right to Be Forgotten)*, 20 COMM. L. & POL'Y 91 (2015); Jeffrey Rosen, *The Right to Be Forgotten*, 64 STAN. L. REV. ONLINE 88 (2012); Michael J. Kelly & David Satola, *The Right to Be Forgotten*, 2017 U. ILL. L. REV. 1 (2017).

data derives from the more technical "internal" functions of intermediaries, which is distinct from the harms resulting from "externally" viewed third party content.[21]

The article proceeds as follows. First, Part II provides a brief discussion on the nature of harmful speech. In particular, this discussion relies on Professor Rodney Smolla's three-part model of harmful speech, which identifies "physical harm, relational harm, and reactive harm."[22] This discussion is important because it establishes an even playing field for analyzing how harm is conceptualized in various parts of the world while also making important distinctions between foreign conceptions of harm and harm as conceived by First Amendment jurisprudence.

Next, Part III conducts a comparative analysis. The analysis draws upon statutes, directives, case law, and secondary scholarship to identify broad-level principles upon which countries' laws of intermediary liability are constructed. The polities whose laws are analyzed are the United States, the European Union, the United Kingdom, Australia, Brazil, India, Japan, South Korea, and South Africa. Countries were selected due to their geographic diversity and position as powerful liberal-democratic polities, and because their populations account for nearly 1.4 billion Internet users as of 2016.[23] Although more than 800 million Internet users live in major global powers such as China and Russia,[24] these countries are excluded due to their autocratic systems of government, reflected (particularly in China) in strict and sophisticated state censorship of much of the social Internet.[25]

21. *See* Orin S. Kerr, *The Problem of Perspective in Internet Law*, 91 GEO. L. J. 357 (2003) (distinguishing two conceptions of the Internet: an internal one involving the technical details that happen behind the scenes, and an external one that users experience in the physical world).

22. SMOLLA, *supra* note 18, at 48. *See infra* notes 28–88.

23. Statistics for population estimates for individual countries (for 2017) come from CIA World Factbook country profiles. *The World Factbook*, CIA (2017), https://www.cia.gov/library/publications/the-world-factbook/; Statistics for percentage of Internet users in individual countries come from: *Individuals Using the Internet: (% of population*, WORLD BANK (2016), https://data.worldbank.org/indicator/IT.NET.USER.ZS?name_desc=false; Statistics for population estimate for the European Union (for 2017) come from official EU statistics. *Living in the EU*, EUROPEAN UNION (2017), https://europa.eu/european-union/about-eu/figures/living_en#tab-1-3; Statistics for percentage of Internet users in the European Union come from official EU statistics. *Internet access and use statistics - households and individuals*, EUROSTAT (2016), https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Internet_access_and_use_statistics_-_households_and_individuals.

Country	Population	Internet %	Internet Users
Australia	23,232,413	88%	20,444,523.44
Brazil	207,353,391	60	124,412,034.6
EU (minus UK)	442,802,756	82	363,098,259.9
India	1,281,935,911	30	384,580,773.3
Japan	126,451,398	92	116,335,286.2
South Africa	54,841,552	54	29,614,438.08
South Korea	51,181,299	93	47,598,608.07
UK	65,648,100	95	62,365,695
USA	326,625,791	76	248,235,601.2
Total	**2,580,072,611**	**74%**	**1,396,685,220**

Estimates of country population and number of Internet users according to 2016-2017 data. (Table compiled by Brett G. Johnson).

24. Statistics for populations estimates for Russia and China come from the CIA World Factbook. CIA, *supra* note 23. Statistics for percentage of Internet users in Russia and China come from WORLD BANK, *supra* note 23.

25. *See* Jonathan Zittrain & John Palfrey, *Internet Filtering: The Politics and Mechanisms of Control*, *in* ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE 29, 33 (Ronald Deibert et al. eds., 2010) (Noting that "China . . . bundles Internet content restrictions with its copyright laws. This set of regulations sets a daunting web of requirements in front of anyone who might access the Internet or provide a service to another Internet user. These rules create a pretext that can be

Finally, Part IV synthesizes the comparative analysis to further identify broad ethical norms by which social media companies should operate. The purpose of this analysis is to compare and contrast legal philosophies for regulating social networking companies so that broad ethical principles may be identified and presented as a way in which these companies might self-regulate. This analysis is important because users, politicians, and jurists are wising to the banes and boons that social media companies bring to society. Social and political pressure is mounting on public officials to do something to mitigate the harms facilitated by digital intermediaries.[26] Meanwhile, digital intermediaries remain an indispensable part of our daily lives, and there is no indication that trend will change anytime soon.[27] It is in the best interest of intermediaries and the public that the former adopt ethical principles with which to self-regulate and mitigate harms that could befall both individuals and democracy.

The goal of this comparative analysis is to arrive at a set of principles upon which social media platforms should act to balance the competing interests of the promotion of political speech, to prevent harms to users, and to ensure the stable financial health of the platforms themselves. A comparative analysis is ideal for achieving this end because the legal contours separating the advancement of one or some of these competing interests over others can be examined in relation to similar and differing contexts. The principles of platform ethics put forth at the end of this analysis are based on an assessment of these contours and which interests they seek to privilege most. Furthermore, these ethical principles can temper the more extreme models of intermediary liability laws, both those that impose a heavy burden on platforms to manage UGC (such as in Brazil and the EU) and the more libertarian model found in the United States. This article concludes with some thoughts on what reforms to these laws might look like based on the principles of platform ethics.

II. HARMS RESULTING FROM SPEECH

The criteria this analysis uses to define harm in the context of intermediary liability laws come from a three-part model of harms that speech can cause, devised by Professor and First Amendment scholar Rodney Smolla: "physical harm, relational harm, and reactive harm."[28] The purpose of using Smolla's model is to lay a foundation for comparative analysis by first discussing how harm is defined within First Amendment jurisprudence. This method does not necessarily mean First

---

used to punish those who exchange undesirable content, even though the law may not be invoked in many instances it might cover."); Emily Parker, *Russia Is Trying to Copy China's Approach to Internet Censorship*, SLATE (Apr. 4, 2017, 1:25 PM), http://www.slate.com/articles/technology/future_tense/2017/04/russia_is_trying_to_copy_china_s_internet_censorship.html.

26. *See* Bill Allison, *Facebook, Google Could Face Tighter Rules on Political Ads*, BLOOMBERG (Feb. 15, 2018, 7:44 PM), https://www.bloomberg.com/news/articles/2018-02-16/facebook-google-could-face-stricter-rules-on-political-ads.

27. Lee Rainie & Janna Anderson, *The Fate of Online Trust in the Next Decade*, PEW RES. CTR. (Aug. 10, 2017), http://www.pewinternet.org/2017/08/10/the-fate-of-online-trust-in-the-next-decade/ ("Many experts say lack of trust will not be a barrier to increased public reliance on the internet. Those who are hopeful that trust will grow expect technical and regulatory change will combat users' concerns about security and privacy. Those who have doubts about progress say people are inured to risk, addicted to convenience and will not be offered alternatives to online interaction.").

28. SMOLLA, *supra* note 18, at 48.

Amendment standards should be the benchmark by which the harms of speech should be judged around the world. Rather, this foundation is useful in its ability to draw baseline principles by which many different definitions and categories of harm can be understood worldwide. Furthermore, these principles can be synthesized to define new types of harm that social media platforms can facilitate, such as the societal-level harms that can arise from the proliferation of so-called "fake news."

A. Physical Harm

United States free speech jurisprudence considers physical harm the worst of the three types of potential harms caused by speech.[29] In the exceptional ethos of free speech in the United States, the capacity of speech to cause physical harm is one area where somewhat clear exceptions have been devised to demarcate when speech falls outside of constitutional protection. Such outlawed speech includes fighting words, incitement to imminent lawless action, and true threats.

The fighting words doctrine comes from the 1942 case *Chaplinsky v. New Hampshire*.[30] In that case, Chaplinsky, a Jehovah's Witness, was arrested and convicted for violating a state breach of peace statute after calling a city marshal "a God damned racketeer and a damned Fascist."[31] The Supreme Court upheld Chaplinsky's conviction, and in so doing crafted the First Amendment exception for fighting words, which the Court defined as words said in another person's face that "by their very utterance inflict injury or tend to incite an immediate breach of the peace."[32]

The incitement to imminent lawless action standard in First Amendment doctrine was refined in the 1969 case *Brandenburg v. Ohio*.[33] That case involved a Ku Klux Klan member, Brandenburg, who was convicted under an Ohio criminal syndicalism law for speaking racist messages to a frenzied crowd.[34] The law prohibited "advocat[ing] . . . the duty, necessity, or propriety of crime, sabotage, violence, or unlawful methods of terrorism as a means of accomplishing industrial or political reform."[35] In a per curiam opinion, the Supreme Court reversed Brandenburg's conviction, holding the following:

> the constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is directed to *inciting or producing imminent lawless action* and is likely to incite or produce such action.[36]

29. David A. Anderson, *Incitement and Tort Law*, 37 WAKE FOREST L. REV. 957, 959 (2002) ("[P]hysical harm . . . seems to present a stronger First Amendment claim than many other types of harm whose ability to trump speech interests is rarely questioned.").
30. 315 U.S. 568 (1942).
31. *Id.* at 569.
32. *Id.* at 572–73.
33. 395 U.S. 444 (1969).
34. *Id.* at 444–45.
35. *Id.* at 445.
36. *Id.* at 447 (emphasis added).

The imminent lawless action standard narrowed the definition of unlawful incitement from the "bad tendency"[37] and "clear and present danger"[38] standards cited by the Court earlier in the twentieth century.

The present doctrinal state of the true threat exception to the First Amendment is somewhat muddled. Generally speaking, United States Courts of Appeal have adopted a "reasonable person" standard for determining whether a threatening statement loses First Amendment protection.[39] However, the United States Supreme Court signaled a possible preference for an alternative test in *Virginia v. Black*, in which the Court held a Virginia statute criminalizing cross burning was unconstitutionally overbroad.[40] Justice O'Connor wrote in a plurality opinion that to convict a person of issuing a true threat, a state must consider whether the "speaker means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals."[41] In a 2015 case involving threatening speech posted to Facebook, the Court sidestepped the doctrinal issue of whether the speaker's intent needed to be taken into account for the speech to be considered a true threat.[42]

In sum, these three doctrines—fighting words, incitement, and true threats—separate the worst of speech-related harms from the body of protected speech in the United States. Physical harm is recognized as the worst possible harm caused by speech because it is direct, immediate, measurable, and often irreparable.[43] As Smolla puts it, "[c]rimes must have victims . . . and the victimization must be palpable, something beyond generalized disgust or disquiet over another's conduct."[44] However, the Court also recognizes speech may indeed best serve its high purpose when it induces a "condition of unrest," creates dissatisfaction with conditions as they are, or even "stirs people to anger,"[45] and therefore such speech needs "breathing space" through expansive legal protections to exist despite the potential physical harms it could trigger.[46]

B. Relational Harm

Relational harm, according to Smolla, involves speech that causes injury to social relationships (e.g., defamation), business relationships (e.g., fraud or false advertising), ownership interests (e.g., copyright), and confidentiality (e.g., leaking

37. Schenck v. United States, 249 U.S. 47 (1919); Debs v. United States, 249 U.S. 211 (1919); Abrams v. United States, 250 U.S. 616, 628 (1919); Gitlow v. New York, 268 U.S. 652 (1925).

38. Cantwell v. Connecticut, 310 U.S. 296 (1940); Terminiello v. City of Chicago, 337 U.S. 1 (1949).

39. *See, e.g.*, Planned Parenthood of Columbia/Willamette, Inc. v. Am. Coal. of Life Activists, 290 F.3d 1058 (9th Cir. 2002); United States v. Bagdasarian, 652 F.3d 1113 (9th Cir. 2011).

40. 538 U.S. 343, 363 (2003).

41. *Id.* at 359 (O'Connor, J., plurality opinion).

42. Elonis v. United States, 135 S. Ct. 2001, 2013 (2015) (Alito, J., dissenting) (lamenting that the Court did not settle the matter of whether the petitioner's speech amounted to a true threat).

43. *See* Clay Calvert, *Hate Speech and Its Harms: A Communication Theory Perspective*, 47 J. COMM. 4, 6–9 (1997) (distinguishing the direct, measurable and immediate nature of physical harms of speech from longer term mental and emotional harms associated with hostile environments created by speech).

44. SMOLLA, *supra* note 18, at 10.

45. Terminiello v. City of Chicago, 337 U.S. 1, 4 (1949).

46. NAACP v. Button, 371 U.S. 415, 433 (1963) (holding that "First Amendment freedoms need 'breathing space' to survive.").

national security secrets).[47] This section focuses only on jurisprudence regarding harms caused by defamation due to the close relationship between defamation and Smolla's third category of reactive harms,[48] and due to the fact that defamation plays a major role in shaping the contours of intermediary liability laws around the world.

The United States Supreme Court constitutionalized defamation law in *New York Times v. Sullivan* by requiring public-official plaintiffs to prove that libelous statements about them were made with "actual malice"—the "knowledge that [the statement] was false or [made] with reckless disregard of whether it was false or not."[49] The same philosophy extends to public figures—those "who are not public officials, but [are] involved in issues in which the public has a justified and important interest,"[50] thereby making them subject to public scrutiny. However, private individuals are generally afforded greater leeway in pursuing defamation suits.[51]

Smolla categorizes defamation as a relational harm due to its close similarity to other business-related harms, such as copyright infringement—it is a harm against property rights.[52] Constitutional scholar Robert Post argues that defamation law in the United States is built on the metaphor that "reputation is capital."[53] Reputation is the fruit "of one's own endeavors."[54] Reputation works hand-in-hand with American capitalism; it can be spent or invested to build up one's fortune, which, in turn, can be invested back into one's good reputation.[55] Post argues that in the United States, the "purpose of the law of defamation is to protect individuals within the market by ensuring that their reputation is not wrongfully deprived of its proper market value."[56]

However, in many cultures, reputation is viewed as an immutable characteristic that is inextricably linked to an individual's sense of honor.[57] Post defines the notion of reputation-as-honor as "a form of reputation in which an individual personally identifies with the normative characteristics of a particular social role and in return personally receives from others the regard and estimation that society accords to that role."[58] This definition of reputation best fits a stratified, hierarchical, or "deference" society rather than a society founded predominantly on market capitalism, such as the United States.[59] Reputation as property is a flexible concept whereby a person can rebuild lost reputation exactly as she would recuperate a lost fortune:

47. SMOLLA, *supra* note 18, at 48.
48. *See infra* notes 61–72.
49. New York Times Co. v. Sullivan, 376 U.S. 254, 280 (1964).
50. Curtis Publ'g Co. v. Butts, 388 U.S. 130, 134 (1967).
51. Gertz v. Welch, 418 U.S. 323, 343 (1974) (holding that "the state interest in compensating injury to the reputation of private individuals requires that a different rule [other than actual malice] should obtain with respect to them.").
52. SMOLLA, *supra* note 18, at 50.
53. Robert C. Post, *The Social Foundations of Defamation Law: Reputation and the Constitution*, 74 CAL. L. REV. 691, 693 (1986).
54. *Id.* at 694 (internal quotations omitted).
55. *Id.* at 695.
56. *Id.*
57. *Id.* at 699. *See also* Brodwyn Fischer, *Slandering Citizens: Insults, Class, and Social Legitimacy in Rio de Janeiro's Criminal Courts*, *in* HONOR, STATUS, AND LAW IN MODERN LATIN AMERICA (Sueann Caulfield, Sarah C. Chambers & Lara Putnam, eds., 2005).
58. Post, *supra* note 53, at 699–700. *See also* Peter F. Carter-Ruck, *Comparative Defamation Law*, 6 INT'L LEGAL PRAC. 3, 6 (1981).
59. Post, *supra* note 53, at 702.

through an entrepreneurial zeal and sound navigation of the market.[60] Honor, however, is a fixed concept, because the "value of honor is the value of a meaningful life."[61] In other words, viewing reputation as honor is more crucially tied to an individual's identity than when reputation is viewed as property. Thus, one can argue the stakes are higher for protecting reputation as honor than they are for protecting reputation as property.

Post's distinct metaphors for reputation show how defamation straddles the line between relational harm and reactive harm when viewed in a global context. In the United States, defamation jurisprudence has moved away from the doctrine of group defamation (particularly against a racial group) that was once outlawed in *Beauharnais v. Illinois*,[62] and toward a more reputation-as-capital conception of defamation.[63] Meanwhile, in other countries, the link between defamation and racism, which Smolla defines as a reactive harm, is much closer. For example, Article 5 of the Brazilian constitution enshrines a right to honor and reputation as well as a right to be free from racism,[64] which Brazilian law views as a crime against honor and dignity.[65]

C. Reactive Harm

Smolla's third category, reactive harm, includes intentional infliction of emotional distress, tortious invasions of privacy, and any type of hate speech.[66] The Supreme Court has raised the standards for plaintiffs suing under the first two categories by imputing the actual malice standard from *Sullivan* into many of these torts, due in large part to their similarity to the tort of defamation.[67] Hate speech has been defined many different ways by many different scholars, but a generic definition for the purposes of this study categorizes hate speech as any speech that attacks and attempts to subordinate any group or class of people, typically spoken by a group with a higher level of social power than the targets of the speech.[68] The targets

60. *Id.* at 695.

61. *Id.* at 701.

62. 343 U.S. 250 (1952).

63. Post, *supra* note 53, at 695.

64. C.F. art. 5 (Braz.) (English version), http://english.tse.jus.br/arquivos/federal-constitution.

65. Lei No. 12.288, de 20 de Julho de 2010, PRESIDÊNCIA DA REPÚBLICA (Braz.) (Statute of Racial Equality), http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2010/Lei/L12288.htm.

66. SMOLLA, *supra* note 18, at 50.

67. Cantrell v. Forest City Publ'g Co., 419 U.S. 245 (1974) (holding that plaintiffs must prove actual malice to successfully recover for the tort of false light invasion of privacy); Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988) (holding that a public figure plaintiff must prove actual malice to successfully recover for the tort of intentional infliction of emotional distress).

68. *See* Calvert, *supra* note 43, at 4 (showing various studies with various definitions of hate speech); Alexander Tsesis, *Dignity and Speech: The Regulation of Hate Speech in a Democracy*, 44 WAKE FOREST L. REV. 497, 504 (2009); Richard Delgado & David H. Yun, *Essay II. Pressure Valves and Bloodied Chickens: An Analysis of Paternalistic Objections to Hate Speech Regulation*, 82 CAL. L. REV. 871, 878-79 (1994); Owen M. Fiss, *The Supreme Court and the Problem of Hate Speech*, 24 CAP. U. L. REV. 281, 290 (1995); Stephanie Farrior, *Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech*, 14 BERKELEY J. INT'L L. 1 (1996); Jean-Marie Kamatali, *The U.S. First Amendment Versus Freedom of Expression in Other Liberal Democracies and How Each Influenced the Development of International Law on Hate Speech*, 36 OHIO N.U. L. REV. 721, 728 (2010); Robert Post, *Hate Speech*, *in* EXTREME SPEECH AND DEMOCRACY 123 (Ivan Hare & James Weinstein eds., 2009) [hereinafter *Hate Speech*]; Tanya Katerí Hernández, *Hate Speech and the Language of Racism in Latin America: A Lens for Reconsidering Global Hate Speech Restrictions and Legislation Models*, 32 U. PA. J. INT'L L. 805, 807–809 (2011).

of such speech typically include racial minorities, women, religious minorities, and the LGBTQ community.[69] Generally, hate speech is only punishable if it contravenes one of the few First Amendment exceptions listed above.[70] According to Smolla, speech that leads to reactive harms deserves the highest level of constitutional protection due to its tendency to implicate public figures or officials, or its tendency to involve important social issues and matters of public concern—factors which greatly outweigh the potential harms of the speech.[71]

However, other scholars view the reactive harms of hate speech in a more nuanced way in an attempt to craft sound First Amendment doctrine for mitigating such harms. Professor Cass Sunstein concedes "the line is sometimes thin between restrictions based on 'harm' and restrictions based on viewpoint of content."[72] However, he holds that the primary factor that should determine whether speech is protected "is whether the speech is a contribution to social deliberation, not whether it has political effects or sources."[73] Thus, Sunstein distinguishes a misogynist tract from pornographic movies, a racist speech to a crowd from face-to-face racial harassment, and a "tract in favor of white supremacy from a racial epithet."[74]

However, Sunstein points out that even within each of those categories, not all hateful words are equal in their potential to cause reactive harm. He writes, "[i]t is obtuseness—a failure of perception or empathetic identification—that would enable someone to say that the word 'fascist' or 'pig' or even 'honky' produces the same feelings as the word 'nigger.'"[75] A deeper moral point can be made from Sunstein's argument: although the many examples of extreme speech listed above receive strong legal protection due to their theoretical social value, their harms are no less real to the people who suffer them.

Nevertheless, in First Amendment jurisprudence, reactive harms are considered the least worrisome of harms caused by speech because they are generally considered less tangible than physical or relational harms.[76] These latter two categories implicate life and property, while reactive harms can be reduced to "hurt feelings."[77]

---

69. *See Hate Speech*, *supra* note 68.

70. *See* R.A.V. v. City of St. Paul, 505 U.S. 377, 391 (1992) (holding that a St. Paul, Minn. ordinance banning symbolic speech (such as cross-burning) that is hateful "on the basis of race, color, creed, religion or gender"—a content-based restriction of speech—was unconstitutionally under-inclusive); Nat'l Socialist Party of Am. v. Vill. of Skokie, 432 U.S. 43 (1977) (holding that delays in issuing parade permits to Nazis were, in and of themselves, a content-based restriction on the Nazi Party's speech); Snyder v. Phelps, 562 U.S. 443 (2011) (holding that allowing an individual, even a private citizen such as Mr. Snyder, to sue for civil damages from emotional distress intentionally inflicted by lawful social speech would lead to a chilling effect on such speech).

71. SMOLLA, *supra* note 18, at 48.

72. CASS R. SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH 174 (1993).

73. Cass R. Sunstein, *Free Speech Now*, 59 U. CHICAGO L. REV. 255, 309 (1992).

74. *Id.*

75. SUNSTEIN, *supra* note 72, at 186.

76. C. Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964, 998 (1978).

77. Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 53–55 (1988):

> [I]n the world of debate about public affairs, many things done with motives that are less than admirable are protected by the First Amendment . . . . 'Outrageousness' in the area of political and social discourse has an inherent subjectiveness about it which would allow a jury to impose liability on the basis of the jurors' tastes or views, or perhaps on the basis of their dislike of a particular expression. An 'outrageousness' standard thus runs afoul of our longstanding refusal to allow damages to be awarded because the speech in question may have an *adverse emotional impact* on the audience.

Certainly, harm to one's mental well-being is nothing trivial. Many scholars who have proposed ideas for stronger regulations against hate speech point out that the damage such speech causes to the well-being of minorities leads to physical (hence, more important) ailments, such as anxiety and depression, which in turn may make minorities retreat from participating in society.[78] However, the inability to create a legal test that would show a "direct causal link"[79] between speech and mental harms (like with true threats, incitement, or actual malice) weighs against Smolla's argument that speech associated with reactive harms often implicates public officials. This makes speech that causes reactive harms the least deserving of an exception from First Amendment protection.

D. Harms and Online Speech in a Global Communication Environment

The three types of harm discussed above reflect the United States' model for categorizing and, in some cases, justifying proscription of harmful speech. In sum, in the United States, unprotected speech must combine a lack of a significant message with the likelihood that some form of harm will befall a targeted recipient.[80] For reactive harms, the task of justifying proscription of speech is difficult.[81] Here, Smolla puts forth his "emotion principle," which claims speech has both emotional and intellectual effects.[82] Under the emotion principle, speech cannot be banned due to its emotional component alone; the intellectual component must be factored in, and even the slightest intellectual value will tip the scale in favor of protecting the speech.[83] Thus, banning speech to prevent harm "may not be satisfied by the outrage or moral opprobrium that a majority of the populace attaches to the activity."[84]

Meanwhile, other countries' legal systems do not package the harms of speech so neatly into physical, commercial, or emotional categories. For instance, the laws of some countries view hateful speech as more closely linked to physical harm than reactive harm. In Brazil, racist speech is criminalized, due in large part to the challenge racism poses to Brazil's founding narrative that the country is a "racial democracy."[85] Similarly, in the European Union, the experience of Nazism and the

---

(emphasis added). *See generally* RESTATEMENT (SECOND) OF TORTS § 46 cmt. j (AM. LAW. INST. 1965) ("[S]ome degree of transient and trivial emotional distress is a part of the price of living among people.").

78. Calvert, *supra* note 43; *see also*, Caroline West, *Words the Silence? Freedom of Expression and Racist Hate Speech*, *in* SPEECH & HARM: CONTROVERSIES OVER FREE SPEECH 222 (Ishani Maitra & Mary Kate McGowan eds., 2012).

79. Clay Calvert, Kara Carnley, Brittany Link & Linda Riedmann, *Conversion Therapy and Free Speech: A Doctrinal and Theoretical First Amendment Analysis*, 20 WM. & MARY J. WOMEN & L. 525, 539 (2014).

80. Frederick Schauer, *Intentions, Conventions, and the First Amendment: The Case of Cross-Burning*, 2003 SUP. CT. REV. 197, 205 (2003).

81. SMOLLA, *supra* note 18, at 51.

82. *Id.* at 46.

83. *Id.*

84. *Id.* at 10.

85. *See* Brett G. Johnson, *Prejudice Against Being Prejudiced: Racist Speech and the Specter of Seditious Libel in Brazil*, 20 COMM. L. & POL'Y 55 (2015) [hereinafter *Prejudice Against Being Prejudiced*].

Holocaust is among the reasons why racist speech is prohibited.[86] These countries also tend to define racist speech in terms of its ability to incite racial hatred and violence, thereby threatening to throw society into chaos.[87]

In the area of relational harm, Professor Robert Post argues that the prevailing interpretation in the United States is that reputation is similar to capital, whereby a bankrupt reputation, just like one's lost fortune, has the potential to be rebuilt.[88] Such an interpretation is one factor behind the exceptional freedoms given to publishers to prevail in defamation lawsuits against public figures.[89] However, Post contends that countries with cultural interpretations of reputation as a reflection of a person's honor tend to have stricter laws governing defamation due to the notion that one's honor is irreplaceable if damaged.[90] In other words, harm to reputation is as irreparable as harm to one's physical self.

Smolla's clean lines distinguishing the harms of speech are becoming further strained, if not outright blurred, due to the new and augmented types of harms perpetuated by speech in our networked communication environment. The Internet's facilitation of anonymous speech has lowered the social cost for speakers to inflict all sorts of harm through their online words.[91] The reach, permanence, and anonymity of Internet communication have the potential to amplify the physical, relational, and reactive harms associated with speech.[92] Law professor Kent Greenawalt identifies four parts to the incitement standard: (1) the extent of the lawlessness of the action the speech is advocating; (2) who the speech is being directed at; (3) the likelihood of the action occurring; and (4) the imminence of the action occurring.[93] Each of these factors provides a layer of protection to speech that has the potential to lead to physical harm.

However, networked communication can allow some types of extreme speech to surpass each of these protective layers. For example, the requirement that the communication be directed immediately at an angry audience may no longer be a sufficient condition for the incited lawless action to be imminent. Professor Lyrissa Lidsky argues that an inflammatory message posted on social media can target both intended and unintended audiences who may be more likely than a restive mob to imminently commit a violent illegal act.[94] Others, however, argue the very idea that almost any controversial or offensive message could be suppressed because of its tendency to incite someone to violence should galvanize society to maintain its

86. *See* Frederick Schauer, *The Exceptional First Amendment*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS 29, 42–43 (Michael Ignatieff ed., 2005) (suggesting that the experience of the Holocaust may explain European legal perspectives toward hate speech).

87. JACOB ROWBOTTOM, *Extreme Speech and the Democratic Functions of the Mass Media*, *in* EXTREME SPEECH AND DEMOCRACY 608, 610 (Ivan Hare & James Weinstein eds., 2009); Hernández, *supra* note 68, at 827; *Prejudice Against Being Prejudiced*, *supra* note 85.

88. Post, *supra* note 53, at 702.

89. *See id.* at 695.

90. *Id.*

91. *See* Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEXAS TECH L. REV. 147, 149 (2011). *See generally*, DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014); Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011); Danielle Keats Citron, *Cyber Civil Rights*, 86 B.U. L. REV. 61 (2009); Yuval Karniel, *Defamation on the Internet—A New Approach to Libel in Cyberspace*, 2 J. INT'L MEDIA & ENTM'T L. 215 (2009).

92. *See* Lidsky, *supra* note 91; *see also* Franks, *supra* note 91, at 228.

93. Kent Greenawalt, *Speech and Crime*, 1980 AM. B. FOUND. RES. J. 645, 653 (1980).

94. Lidsky, *supra* note 91, at 149.

standard of only outlawing the rare case of directly inciting imminent lawless action, online or off.[95] The issue of terrorist propaganda presents a thorny dilemma for this debate. At its very core, such content could be considered political speech—no different than the speech of other extremist groups advocating for death, destruction, and the overthrow of government. Yet such speech could also prove very successful in recruiting disaffected youths to become terrorists, thereby indirectly inciting violence.

Meanwhile, networked communication has seen the generation of new categories of harmful speech, such as "revenge porn," which involves posting nude images online of an ex-romantic partner out of spite, and "cyber-harassment," which involves persistently inflicting substantial emotional distress against an individual through online communications.[96] Indeed, in a networked communication environment, the reactive harms associated with hate speech have the potential to morph into physical harms when they take the form of cyber-harassment or abuse of an individual.[97] Despite this rise in the level of harm, few if any legal options exist for targets of such invectives, shifting the focus from pie-in-the-sky legal remedies to calling on digital intermediaries to mitigate these harms through self-regulation and tech-based fixes.[98]

It is important to unpack this dilemma as it is central to the analysis of this article. Digital intermediaries must abide by the laws of other countries and the various ways in which they define the harms of speech, both in general and in an online context.[99] So, how do such platforms balance the goal of self-regulation with following the many laws set out before them? Scholars Jack Goldsmith and Tim Wu have suggested the best way to achieve this balance is to create a bordered Internet where unlawful content can be policed automatically using geolocation.[100] However, such a policy neglects the harms of online speech in a country like the United States where, as will be discussed next, protections against intermediary liability are very strong. Moreover, it raises the specter of whether lawful and significant political speech in certain parts of the world could get caught up in the dragnet of platforms' automatic policing of allegedly harmful speech according to geolocation.[101] To properly conceptualize "platform ethics" and consider the ways in which intermediaries can balance self-regulation with legal compliance, an understanding of various models of intermediary liability laws is necessary.

---

95. *See* Lynn Adelman & Jon Deitrich, *Extremist Speech and the Internet: The Continuing Importance of* Brandenberg, 4 HARV. L. & POL'Y REV. 361, 370 (2010); *see also* L.A. Powe, Jr., Brandenburg*: Then and Now*, 44 TEXAS TECH L. REV. 69 (2011).

96. *See generally* CITRON, *supra* note 91.

97. *Id.* at 69; Franks, *supra* note 91, at 246.

98. *See* Eric E. Schmidt, *Eric Schmidt on How to Build a Better Web*, N.Y. TIMES (Dec. 7, 2015), http://www.nytimes.com/2015/12/07/opinion/eric-schmidt-on-how-to-build-a-better-web.html (calling on online intermediaries such as Google—the company he cofounded—to create "spell-checkers . . . for hate and harassment.").

99. *See, e.g.*, Ronald Deibert & Rafal Rohozinski, *Beyond Denial: Introducing Next-Generation Information Access Controls*, *in* ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE 3 (Ronald Deibert et al. eds., 2010); LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE 168 (2014).

100. JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 10 (2006).

101. *See infra* Part III.A.

III. COMPARATIVE ANALYSIS

Understanding the philosophies behind laws of intermediary liability can enrich the discussion of the potential ethical obligations that intermediaries have when it comes to dealing with harmful UGC. In particular, one must understand how liability models differ in terms of privileging self-regulation within the industry and protecting the consumers who use social media platforms. Indeed, the approaches of the following nine polities vary rather widely on this spectrum. Furthermore, none are completely static, with several having made major—and not necessarily clear—shifts in their approach to intermediary liability in recent years.

The preceding discussion of conceptions of harmful speech, particularly in the context of networked communication, will prove useful in a comparative analysis of various legal regimes defining intermediary liability in relation to harmful third-party speech. Physical, relational, and reactive harms are regarded with varying degrees of severity in the laws of various countries. Social media platforms complicate the already messy distinctions between these types of harms because, as discussed above, they can both amplify and blur the lines between these harms. Therefore, the following comparative analysis should be viewed through the lens of how laws of intermediary liability view the shifting sands of the harms of speech on social media.[102]

A. United States Model

The United States' approach to intermediary liability is enshrined in § 230 of the 1996 Communications Decency Act ("CDA").[103] The law grants digital intermediaries (which the law refers to as "interactive computer services") immunity from civil liability for content published by third parties on its platforms even when they are notified of the presence of the content or when they choose to take control over the content and remove it in a "Good Samaritan" act.[104] Knowledge of other types of tortious material does not force intermediaries to remove the material.[105] In the preamble to § 230, Congress declared that digital intermediaries deserve "a minimum of government regulation" because they "offer users a great degree of control over the information that they receive, as well as the potential for even greater control in the future as technology develops."[106] It called the Internet a "forum for a true diversity of political discourse," facilitated by digital intermediaries.[107] Congress declared that its intent in passing the law was "to preserve the vibrant and

102. It should be known that these laws apply to intermediaries that are global in scope—such as Facebook, Twitter, and YouTube—as well as autochthonous platforms. A few examples of the latter type will be listed *infra* in each polity's respective subsection, when applicable.

103. 47 U.S.C. § 230 (2018).

104. 47 U.S.C. § 230(c)(2)(A) ("No provider or user of an interactive computer service shall be held liable on account of . . . any action voluntarily taken in good faith to restrict access" to tortious third-party content).

105. *See* Zeran v. AOL, Inc., 129 F.3d 327, 331 (4th Cir. 1997) (holding that § 230 gave AOL immunity from liability for defamatory third-party content despite the fact that it had been made aware of the existence of the content and had taken steps to remove it).

106. 47 U.S.C. § 230(a)(2), (4).

107. *Id.* § 230(a)(3).

competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation."[108]

Section 230 is based on two important rationales related to the connection between intermediaries, commerce, and freedom of expression. First, it seeks to protect intermediaries from having to incur the great costs necessary to sift through the terabytes of content they host and weed out defamatory material, because this could potentially chill their desire to host otherwise free expression online.[109] Second, § 230 seeks to prevent individuals (whether public figures or not) from having an incentive to serve Internet Service Providers ("ISPs") with potentially successful requests for taking down content, thus protecting speech from frivolous takedowns.[110]

United States case law involving § 230 reinforces Congress's philosophy on the broad, speech-friendly benefits of the provision despite its obvious side effect of allowing potentially harmful speech to flourish. In *Zeran v. AOL*, the Fourth Circuit stated that § 230, quite simply, was a "policy choice . . . not to deter harmful speech through the separate route of imposing tort liability on companies that serve as intermediaries," thereby "maintain[ing] the robust nature of Internet communication."[111] In *DiMeo v. Max*, the United States District Court for the Eastern District of Pennsylvania averred that "we should expect such [harmful] speech to occur in a medium in which citizens from all walks of life have a voice."[112]

Courts have held that digital intermediaries lose their immunity from liability under § 230 if they "materially contribute" to the creation of unlawful content on their platforms. For example, in *Fair Housing Council of San Fernando Valley v. Roommates.com*, the Ninth Circuit held that a dropdown menu allowing users to select the race and gender of potential roommates sought through Roommates.com materially contributed to the violation of federal Fair Housing Act.[113] However, the Sixth Circuit held that an employee of a website goading users into posting defamatory statements on the site did not amount to a material contribution to the creation of that content.[114] Together, these cases highlight the power of § 230 in affording exceptional immunity from liability to platforms.

B. European Model

The goal of focusing on intermediary liability at the supranational level across the European Union ("EU") is to allow for key principles and values surrounding intermediary liability across the continent to be more easily identified.[115] Intermediary liability laws of each EU member state will not be the focus of this section.

---

108. *Id.* § 230(b)(2).

109. *See* 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 (2010).

110. *See Zeran*, 129 F.3d at 333 (arguing that "notice-based liability for interactive computer service providers would provide third parties with a no-cost means to create the basis for future lawsuits.").

111. *Id.* at 330–31.

112. 433 F. Supp. 2d 523, 533 (E.D. Pa. 2006).

113. 521 F.3d 1157 (9th Cir. 2008).

114. Jones v. Dirty World Entm't Recordings LLC, 755 F.3d 398 (6th Cir. 2014).

115. According to estimates from StatCounter, U.S. based social media companies (Facebook, Twitter, etc.) account for more than 99% of all social media use in the EU. *See Social Media Stats Europe Sept*

However, attention will be paid to a 2017 German law imposing hefty fines against social media platforms for unlawful third-party content.

The regime of intermediary liability in the EU is based on Directive 2000/31/EC—the so-called "e-Commerce Directive."[116] Recital 46 of the e-Commerce Directive states digital intermediaries—here referred to as providers of an "information society service"—benefit from a limitation of liability if "upon obtaining actual knowledge or awareness of illegal activities[, they] act expeditiously to remove or to disable access to the information concerned."[117] Recital 48 says EU Member States may "apply duties of care" on digital intermediaries, "which can reasonably be expected from them and which are specified by national law, in order to detect and prevent certain types of illegal activities."[118]

Various European courts have broadly interpreted "illegal activities" under this Directive to include content that causes relational or emotional harms,[119] such as defamation, violation of privacy, and hate speech.[120] A duty of care is established when the alleged victim of such harms appropriately notifies the digital intermediaries of the content in question on their platforms, thereby putting these companies on the legal hook for removing it.[121] However, this duty of care is only established in a notice-and-takedown regime.[122] Article 15(1) of the e-Commerce Directive prohibits Member States from "impos[ing] a general obligation on providers . . . to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity."[123] Therefore, European officials can only call on digital intermediaries to follow a *moral* duty to police harmful speech published on their platforms, as they did following the 2015 *Charlie Hebdo* attacks.[124] Meanwhile, the EU's 2017 Terrorism Directive decrees that "an effective means of combating terrorism on the Internet is to remove online content

---

2017 - Sept 2018, STATCOUNTER, http://gs.statcounter.com/social-media-stats/all/Europe (last visited Aug. 26, 2018).

116. Directive 2000/31, of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, 2000 O.J. (L 178) 1, 14 (EC).

117. *Id.* at (46).

118. *Id.* at (48).

119. SMOLLA, *supra* note 18, at 48–49.

120. *See* Timothy Pinto, Niri Shan, Stefan Freytag, Elisabeth von Braunscheig & Velérie Aumage, *Liability of Online Publishers for User Generated Content: A European Perspective*, 27 COMM. LAW. 5 (2010).

121. Marcelo Thompson, *Beyond Gatekeeping: The Normative Responsibility of Internet Intermediaries*, 18 VAND. J. ENT. & TECH. L. 783, 806 (2016).

122. *Id.*

123. Directive 2000/31, art. 15, of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market, 2000 O.J. (L 178) 1 (EC).

124. Joint Statement of the Ministers of Interior, European Union (Jan. 11, 2015), https://eu2015.lv/images/news/2015_01_11_Joint_statement_of_ministers_for_interrior.pdf. The statement read, in part:

> We are concerned at the increasingly frequent use of the Internet to fuel hatred and violence and signal our determination to ensure that the Internet is not abused to this end, while safeguarding that it remains, in scrupulous observance of fundamental freedoms, a forum for free expression, in full respect of the law. With this in mind, the partnership of the major Internet providers is essential to create the conditions of a swift reporting of material that aims to incite hatred and terror and the condition of its removing, where appropriate/possible.

constituting a public provocation to commit a terrorist offence [*sic*] at its source."[125] The Terrorism Directive stipulates the following:

> Member States shall take the necessary measures to ensure the prompt removal of online content constituting a public provocation to commit a terrorist offence [*sic*] . . . that is hosted in their territory. They shall also endeavour [*sic*] to obtain the removal of such content hosted outside their territory.[126]

In keeping with the eCommerce Directive, the Terrorism Directive does not require intermediaries to regularly monitor their platforms for terrorist content.[127]

The goal of the EU model of intermediary liability, as defined by the laws above, is to incentivize self-regulation by digital intermediaries by encouraging a proactive approach whereby companies would actively screen user-generated content and remove the manifestly unlawful material before upset users have a chance to notify them and thus place them within the prospects of liability.[128] Thus, "Only when contents [are] manifestly unlawful—so that intermediaries would not have to appreciate their lawfulness—would the latter be required to react and eventually take them down or restrict access to them."[129] As with § 230, the philosophy behind the European approach to intermediary liability is that "private regulation is less dangerous than public regulation when it comes to the defence [*sic*] of freedom of expression."[130]

However, a 2015 ruling by the European Court of Human Rights ("ECtHR") appeared to go against the eCommerce Directive's policy of prohibiting general monitoring by intermediaries, throwing the state of intermediary liability law in the EU into disarray.[131] In *Delfi AS v. Estonia*, a divided Grand Chamber of the ECtHR held that Delfi, an intermediary that maintained a comment section for news articles posted on its site, could be held liable for comments that were defamatory, amounted to hate speech, or incited violence because the intermediary was made aware and it commercially benefited from the comments.[132] The court argued that an intermediary's "economic interest in the publication of comments" is no different from a publisher of print material.[133]

---

125. Directive 2017/541, of the European Parliament and of the Council of 15 March 2017 on Combatting Terrorism and Replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, 2017 O.J. (L 88) 22 (EU).

126. *Id.* art. 21 § 1.

127. *Id.* at (23).

128. Sophie Stalla-Bourdillon, *Sometimes One is Not Enough! Securing Freedom of Expression, Encouraging Private Regulation, or Subsidizing Internet Intermediaries or All Three at the Same Time: The Dilemma of Internet Intermediaries' Liability*, 7 J. INT'L COM. L. & TECH. 154, 164–65 (2012).

129. *Id.* at 162.

130. *Id.* at 164.

131. *See* Lisl Brunner, *The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after* Delfi v. Estonia, 16 HUM. RTS. L. REV. 163, 164 (2016); Bart van der Sloot, *The Practical and Theoretical Problems with 'Balancing'* Delfi, Coty *and the Redundancy of the Human Rights Framework*, 23 MAASTRICHT J. EUR. & COMP. L. 439, 448 (2016).

132. Delfi AS v. Estonia, App. No. 64569/09, 2015 Eur. Ct. H.R. 60–61, https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-155105%22]}.

133. *Id.* at 44.

Furthermore, the court held that the societal interests in stanching the flow of hate speech online was justification for holding intermediaries like Delfi liable for comments posted to their platforms. In particular, the court wrote the following:

> [I]n cases such as the present one, where third-party user comments are in the form of hate speech and direct threats to the physical integrity of individuals . . . the rights and interests of others and of society as a whole may entitle Contracting States to impose liability on Internet news portals, without contravening Article 10 of the [European] Convention [on Human Rights protecting the right to freedom of expression], if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties.[134]

However, the court was not unanimous in this conclusion that intermediaries like Delfi should have such a heavy burden on policing hate speech. Two judges dissented in the *Delfi* case and argued that because the majority's opinion requires intermediaries to decide between monitoring user comments for racist or defamatory speech or not allowing comments at all, intermediaries will choose the latter, thus giving them "an invitation to self-censorship at its worst."[135] Although the dissenting judges acknowledged the comments were racist and defamatory and the potential to incite violence were reason enough to prosecute the individuals who posted them, they argued that Delfi should not be held liable merely for opening up a discussion forum about issues of public concern and having that forum commandeered by others to express unlawful opinions.[136]

Meanwhile, in June 2017, Germany passed the Act to Improve the Enforcement of Rights on Social Networks, also known as the "Network Enforcement Act."[137] The law, which was enacted in October 2017, requires social media networks with more than two million users to remove UGC that is "clearly illegal" within 24 hours of receiving a notice from a user about the content.[138] If the content is not clearly illegal, the social network is given seven days to investigate and decide whether or not it is worthy of deletion.[139] The law's list of illegal content includes propaganda of unconstitutional organizations, encouragement of violent crimes, and incitement to hatred.[140] The law mandates any social networks that receive more than 100 notifications of infringing content in a year must publish biannual reports regarding how they handle those notifications.[141] A social network that intentionally or negli-

134. *Id.* at 59–60.
135. *Id.* at 68 (Sajó, J., & Tsotsoria, J., dissenting).
136. *Id.* at 72–73, 77.
137. *See generally* Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], June 30, 2017, Federal Law Gazette at 3352, (Ger.), https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html. For a concise summary of the law in English, *see* Jenny Gesley, *Germany: Social Media Platforms to Be Held Accountable for Hosted Content Under "Facebook Act"*, LIBR. OF CONGRESS (July 11, 2017), https://www.loc.gov/law/foreign-news/article/germany-social-media-platforms-to-be-held-accountable-for-hosted-content-under-facebook-act/.
138. Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], June 30, 2017, Federal Law Gazette at 3353, art. 1 § 3 ¶ 2 no. 2 (Ger.), https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html.
139. *Id.* art. 1 § 3 ¶ 2 no. 3.
140. *Id.* art. 1 § 1 ¶ 3.
141. *Id.* art. 1 § 3 ¶ 1.

gently fails to follow the mandates of the law faces a fine of up to 50 million euros.[142] Coupled with the *Delfi* decision, Germany's Network Enforcement Act indicates a growing trend in EU law toward strict regulation of digital intermediaries.

C. United Kingdom Model

As the United Kingdom ("UK") continues its slow yet inevitable exit from the EU, it is important to study how UK law treats intermediary liability separately from the EU.[143] The UK is also worth studying due to how active the British government has been in addressing issues related to intermediary liability and mitigating harm since 2013, and particularly since the phenomenon of fake news became an issue in 2016.[144] In January 2018, the UK Department for Digital, Culture, Media, and Sport published a two-page Digital Charter (the "Charter") putting forth policy goals for regulating online intermediaries by establishing "norms and rules for the online world."[145]

The Charter lists three priorities: (1) "protecting people from harmful content and behaviour [*sic*], including building understanding and resilience, and working with industry to encourage the development of technological solutions"; (2) "looking at the legal liability that online platforms have for the content shared on their sites, including considering how we could get more effective action through better use of the existing legal frameworks and definitions"; and (3) "limiting the spread and impact of disinformation intended to mislead for political, personal and/or financial gain."[146] The Charter called on a multi-stakeholder approach for addressing these priorities that involved self-regulation within the tech industry alongside government regulation.[147]

Britain is also home to the UK Internet Service Providers' Association ("ISPA UK"), a trade organization comprising Britain's Internet service providers.[148] Its Code of Practice explicitly excludes issues related to third-party content.[149] For example, the UK Code encourages members to "use their reasonable endeavours [*sic*]

142. *Id.* art. 1 § 4 ¶ 5.

143. According to estimates from StatCounter, U.S. based social media companies (Facebook, Twitter, etc.) account for more than 99% of all social media use in the UK. *See Social Media Stats United Kingdom - September 2018*, STATCOUNTER, http://gs.statcounter.com/social-media-stats/all/united-kingdom (last visited Oct. 7, 2018); *see also Market Share Held by the Leading Social Networks in the United Kingdom (UK) as of July 2018*, STATISTA, https://www.statista.com/statistics/280295/market-share-held-by-the-leading-social-networks-in-the-united-kingdom-uk/ (last visited Oct. 7, 2018).

144. *See generally Intermediary Liability*, MEDIA POL'Y PROJECT BLOG, http://blogs.lse.ac.uk/mediapolicyproject/tag/intermediary-liability/ (last visited Oct. 7, 2018); *see also Closing Date: UK 'Fake News' Inquiry*, MEDIA POL'Y PROJECT BLOG (Mar. 3, 2017), http://blogs.lse.ac.uk/mediapolicyproject/event/uk-fake-news-inquiry-closes/ (discussing an inquiry launched by the UK Parliament's Culture, Media and Sport Committee into defining fake news and discussing potential policy solutions for abating it); Brett G. Johnson, *British PM Calls for Nationwide Default Filters to Combat Internet Pornography*, 18 SILHA CTR. STUDY MEDIA ETHICS & L. 30 (2013) [hereinafter *British PM*].

145. UK DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT, POLICY PAPER: DIGITAL CHARTER (2018) (UK), https://www.gov.uk/government/publications/digital-charter/digital-charter.

146. *Id.*

147. *Id.*

148. *See generally* ISPA UK, https://www.ispa.org.uk (last visited Oct. 7, 2018).

149. *ISPA Code of Practice*, INTERNET SERVICE PROVIDERS' ASS'N (Apr. 19, 2002) (UK), napod.org.uk/ispa_code_of_practice.doc.

to ensure th[at] . . . Services (*excluding* Third Party Content) and Promotional Material do not contain . . . material inciting violence, cruelty or racial hatred,"[150] and "are not used to promote or facilitate practices which are contrary to UK law."[151] Meanwhile, the only concrete legislation governing intermediary liability in the UK deals with issues of libel as defined by the 2013 Defamation Act—part of a broad attempt to reform the UK's libel laws.[152] Section 5 of the Defamation Act stipulates that a digital intermediary is not liable if it can prove that it was not responsible for the posting of defamatory statements on its platform.[153] However, if the defamed party notifies the intermediary of the defamatory statements and asks it to remove them, the intermediary could be held liable for the statements if it fails to remove them.[154] Thus, the UK approach to intermediary liability is similar to the United States' approach in that it seeks to privilege self-regulation within the industry rather than state regulation.

D. South African Model

In South Africa,[155] the Electronic Communications and Transactions Act of 2002 establishes a notice-and-takedown regime similar to other models discussed above.[156] Intermediaries are required to remove infringing content upon receiving actual knowledge of it,[157] but they have no general obligation to monitor content on their platforms.[158] The law also permits the Director General of the South African Department of Communications to appoint a "cyber inspector" who has the power to monitor websites for illegal activity and issue takedown notices.[159]

The Internet Service Providers Association ("ISPA") of South Africa, a trade organization representing all Internet service providers and intermediaries established in South Africa (though not necessarily all those operating in the country), maintains a Code of Conduct that sets voluntary self-regulatory principles that ensure intermediaries not only comply with the law, but that they do so transparently.[160] The Code of Conduct, which was last updated in 2016, states that "[t]here is no general obligation on any ISPA member to monitor services provided to customers, but a member is obliged to take appropriate action where it becomes aware of any unlawful content or conduct."[161] The Code of Conduct obliges members to

150. *Id.* § 2.2.1 (emphasis added).
151. *Id.* § 2.2.2.
152. Defamation Act 2013, c. 26, §§ 5, 10 (Eng.).
153. *Id.* § 5(2).
154. *Id.* § 5(3)(b)-(c).
155. U.S. based social networks dominate in South Africa. *See Penetration of Leading Social Networks in South Africa as of 3rd Quarter 2017*, STATISTA, https://www.statista.com/statistics/284468/south-africa-social-network-penetration/ (last visited Oct. 7, 2018). However, the network 2go, which was founded in Johannesburg and is currently based out of Cape Town, is a popular autochthonous social network. *See* 2GO, http://www.2go.im/.
156. Electronic Communications and Transactions Act 25 of 2002 § 75, (S. Afr.), http://www.internet.org.za/ect_act.html.
157. *Id.*
158. *Id.* § 78(1).
159. *Id.* §§ 80, 81.
160. *See Code of Conduct*, INTERNET SERVICE PROVIDERS' ASS'N (June 1, 2016) (S. Afr.), https://ispa.org.za/code-of-conduct/.
161. *Id.* § J(26).

establish clear guidelines for how users can initiate notice-and-takedown procedures, keep records of all requests for removing content, and provide regular reports on such requests to the ISPA.[162]

By working in concert with South African law governing intermediaries, the Code of Conduct is designed to hold intermediaries accountable for their actions by ensuring that they do not remove too much or too little content. In doing so, the Code of Conduct offers a means for intermediaries to balance the competing goals of promoting free expression and preventing the proliferation of harmful UGC. By governing ISPs' practices toward handling third-party content, the South African ISPA is distinct from the ISPA UK, which, as noted above, explicitly excludes issues related to third-party content from its Code of Conduct.[163] However, it is similar to the UK approach to intermediary liability in that it seeks to incentivize self-regulation over state regulation.

E. Australian Model

Outside of the realm of copyright law, Australia has no specific legislation dealing with intermediary liability in the harmful third-party context.[164] Thus, judge-made law has defined the contours of intermediary liability in Australia, and courts have been inconsistent on recent issues. One example of conflicting common law is on the issue of whether Google is considered a publisher of, and therefore considered liable for, defamatory search results. In *Trkulja v. Google*, the Supreme Court of Victoria held that Google is considered a publisher in the context of search results.[165] In particular, the court held that because employees of Google possess "skill and expertise . . . employed by Google for the purpose of creating a search engine," and because "Google intends its search engines to publish material on the [I]nternet in response to user queries," Google must be considered a publisher.[166] The court called Google's argument an attempt to "confer immunity out of thin air,"[167] and held "[i]f Google is to have immunity from suit, it must be bestowed upon it by the legislature."[168]

However, the Supreme Court of New South Wales held in a separate case that Google is not considered a publisher of defamatory search results because the search results are generated via algorithm, not human activity.[169] Like the Victoria Supreme Court in *Trkulja*, the New South Wales court acknowledged algorithms were created by humans, but the latter disagreed that this connection rose to the level of human activity required for Google to be considered a publisher.[170] The New South Wales court noted that because the individual notified Google of the

162. *Id.* §§ J(29)-(31).
163. *ISPA Code of Practice*, INTERNET SERVICE PROVIDERS' ASS'N (Apr. 19, 2002) (UK), napod.org.uk/ispa_code_of_practice.doc.
164. U.S. based social media firms dominate the Australian market, accounting for more than 99% of the country's social media use. *See generally* David Cowling, *Social Media Statistics Australia – April 2018*, SOCIALMEDIANEWS (May 1, 2018), https://www.socialmedianews.com.au/social-media-statistics-australia-april-2018/.
165. *Trkulja v Google Inc.* [2015] VSC 635 (17 November 2015) 67 (Austl.).
166. *Id.* at 54.
167. *Id.* at 77.
168. *Id.* at 75.
169. *Bleyer v Google Inc.*, [2014] NSWSC 897, 83-85 (12 August 2014) (Austl.).
170. *Id.* at 83.

posts and asked to have them removed, that might turn Google into a publisher under Australian law, thereby making it liable for the posts.[171] However, the court held that because the plaintiff claimed only three people in Australia saw the defamatory search results, the cost of holding Google responsible for the search results outweighed any benefits from mitigating any harms in doing so—in other words, the case was dismissed for lack of proportionality.[172]

It is illustrative to contrast both of these cases with a case involving defamatory statements made on a website with a much smaller operation than Google's. In *Piscioneri v. Brisciani*, the Supreme Court of the Australian Capital Territory held that an individual who ran a website that contained defamatory posts from users, and who encouraged users to make those posts, must be considered a publisher and thus liable for the posts of the users on his site.[173] Thus, according to Australian case law, the line distinguishing the terms of liability for global intermediaries such as Google and local small-time intermediaries appears to be faded, gray, and porous.

F. Brazilian Model

In Brazil,[174] intermediary liability is codified in a 2014 law known as the "Marco Civil da Internet" (the "Marco Civil").[175] Roughly translated as an "Internet Bill of Rights," the statute establishes, among other things, that Brazilian citizens have a right to net neutrality.[176] The Marco Civil lists a special provision under Article 21 for the phenomenon of "revenge porn," whereby intermediaries will be held criminally liable if they either purposefully host or fail to remove revenge porn photos on their platforms.[177] Articles 18 and 19 of the Marco Civil provide that digital intermediaries are immune from civil liability for third-party content unless they fail to remove defamatory or racist content after receiving a valid court order asking them to do so.[178] The connection between defamation and racism in this law should not be overlooked.

Section 3 of Article 140 in Brazil's Penal Code states the following: "If [a] defamatory act involves the use of references to race, color, ethnicity, religion, origin, elderly status[,] or disability" the penalty will be "imprisonment of one to three years and a fine."[179] A 1997 anti-racism law reiterates Brazil's intention to

171. *Id.* at 85–87.
172. *Id.* at 62.
173. *Brisciani v Piscioneri [No. 4]* (2016) ACTCA 32, 17 (Austl.), 2016 WL 4239922. The facts in this case are analogous to those in *Jones v. Dirty World Entertainment Recording, LLC*. *See* Jones v. Dirty World Entm't Recordings, LLC, 755 F.3d 398, 415–16 (6th Cir. 2014).
174. U.S. based WhatsApp and Facebook dominate the Brazilian social media market. *See Brazil: Most Popular Social Network Apps as of June 2017*, STATISTA, https://www.statista.com/statistics/746969/most-popular-social-networkapps-brazil/ (last visited Oct. 7, 2018).
175. Lei No. 12.965, de 23 de Abril de 2014, CÓDIGO CIVIL [C.C.] (Braz.). *See* Alexandre Pontieri, *Título: Marco Civil da Internet - Neutralidade de Rede e Liberdade de Expressão*, JUS.COM.BR (July 2018), https://jus.com.br/artigos/67822/titulo-marco-civil-da-internet-neutralidade-de-rede-e-liberdade-de-expressao.
176. Lei No. 12.965 art. 3 § IV.
177. *Id.* art. 21.
178. *Id.* art. 18, 19.
179. Lei No. 2.848, de 7 de Dezembro de 1940, CÓDIGO PENAL [C.P.] (Braz.).

enforce racism as a defamatory crime, or a "crime against honor."[180] Indeed, Brazilian law views racism as anathema to the identity of the Brazilian nation, and thus punishes racism under the philosophy that it is akin to seditious libel.[181]

One particular example of action taken against a digital intermediary (albeit prior to the passing of the Marco Civil) is illustrative of the Brazilian context of intermediary liability. In September 2012, two videos appeared on YouTube alleging Alcides Bernal—mayoral candidate for the city of Campo Grande in the Brazilian state of Mato Grosso do Sul—hated poor people, unlawfully enriched himself, paid an ex-lover to abort a child he fathered, denied being the child's father after he was born, and then beat the child after finally admitting he was the father.[182] Bernal filed a lawsuit against Google Brazil, the owner and operator of YouTube in Brazil, for publishing defamatory electoral propaganda against him in the run-up to an election, which is a violation of Article 243 of the Brazilian Electoral Code.[183]

Wanting to uphold the rules for conducting free and fair elections as painstakingly defined in the Electoral Code, Judge Flávio Saad Perón of the 35th Electoral Zone of the municipality of Campo Grande—a division of the Regional Electoral Court, known as the Tribunal Regional Eleitoral of the state of Mato Grosso do Sul—ordered the video be taken down.[184] However, the head of Google Brazil, Fabio José Silva Coelho, refused to obey the order, citing a commitment to upholding the values of free speech.[185] Judge Saad then ordered Coelho placed under house arrest for disobeying a judge's order—a violation of Article 347 of the Electoral Code—and ordered a 24-hour suspension of *all* Google and YouTube services in the state.[186] The judge's order attracted national and international media attention on the otherwise ordinary and relatively insignificant election.[187] Google Brazil released a statement saying it was "appealing the decision that ordered the removal of the YouTube video because, in being a platform, Google is not responsible for

180. Hernández, *supra* note 68, at 828. *See* Lei No. 9.459, de 13 de Maio de 1997, CÓDIGO CIVIL [C.C.] (Braz.).

181. *See Prejudice Against Being Prejudiced*, *supra* note 85, at 57.

182. Bryan Bishop, *Google Complies with Brazilian Court Order to Pull Political Video from YouTube*, VERGE (Sept. 27, 2012, 9:06 PM), https://www.theverge.com/2012/9/27/3420574/google-youtube-brazil-court-order-pulls-political-video; *see also* Brad Hayes, *Google Exec Questioned Over Brazil Election Video*, REUTERS (Sept. 26, 2012, 4:45 PM), https://www.reuters.com/article/net-us-google-brazil/google-exec-questioned-over-brazil-election-video-idUSBRE88P1OX20120926?feedType=RSS&feedName=technologyNews&utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+reuters%2FtechnologyNews+%28Reuters+Technology+News%29.

183. Felipe Correa, *Brazil Confronts Google – and it's Personal*, FREE SPEECH DEBATE (Nov. 1, 2012), http://freespeechdebate.com/case/brazil-confronts-google-and-its-personal/. *See* Lei No. 4.737, de 15 de Julho de 1965, CÓDIGO CIVIL [C.C.] art. 243, de 09.04.1964 (Braz.).

184. Bishop, *supra* note 182.

185. Reuters, *Top Google Executive in Brazil Faces Arrest over Video*, N.Y. TIMES (Sept. 25, 2012), https://www.nytimes.com/2012/09/26/business/global/top-google-executive-in-brazil-faces-arrest-over-video.html.

186. Adi Robertson, *Brazilian Judge Orders Arrest of Local Google President Over Negative Election Videos on YouTube*, VERGE (Sept. 25, 2012, 2:09 PM), https://www.theverge.com/2012/9/25/3406238/brazil-google-president-arrest. *See* Lei No. 4.737, de 15 de Julho de 1965, CÓDIGO CIVIL [C.C.] art. 347, de 09.04.1964 (Braz.).

187. *Brazilian Police Detain Local Google President*, BBC (Sept. 27, 2012), http://www.bbc.com/news/world-latin-america-19737364.

the content posted on its site."[188] The company did not comment on Coelho's arrest. However, on September 26, 2012, Google removed the videos from YouTube.[189]

The philosophical foundation of the Brazilian approach to intermediary liability is based on the Brazilian Consumer Protection Code ("CPC") of 1990,[190] which lays out numerous rights of consumer protection. The philosophical thrust of the CPC is that consumers deserve protection from businesses because, ultimately, consumers are the reason businesses are in business to begin with; in other words, consumers deserve a substantial amount of legal power over the businesses that profit off of them.[191] In the context of intermediary liability, the theory is that because online communication platforms profit off of users by commodifying both their content and their data, these platforms should ultimately respond to users when this venture turns harmful.[192] This philosophy parallels the reasoning in the ECtHR's *Delfi* decision that the commercial incentive of intermediaries to profit off of harmful speech demands a legal remedy to de-incentivize the facilitation of such speech.[193] Therefore, the Brazilian approach to intermediary liability can be viewed as heavily favoring consumer protection over affording immunity to platforms, even in spite of the fact that significant political speech may be chilled to achieve that end.

G. Indian Model

In India,[194] Article 79 of the country's Information Technology ("IT") Act of 2008 stipulates that a digital intermediary is not held liable for third-party content except when it either materially contributes to the creation of the content, or if it receives actual knowledge that the content is unlawful.[195] In 2011, the Indian government published the "Information Technology (Intermediary guidelines) Rules" in its official gazette to further define what might make third-party content unlawful.[196] This includes content that "is grossly harmful, harassing, blasphemous, defamatory, obscene, pornographic, paedophilic [*sic*], libelous, invasive of another's privacy, hateful, or racially [or] ethnically objectionable, disparaging, relating or encouraging money laundering or gambling, or otherwise unlawful in any manner whatever."[197]

188. Megan Geuss, *Google's Brazil Chief Detained by Federal Police Over YouTube Video*, ARS TECHNICA (Sept. 26, 2012, 9:20 PM), https://arstechnica.com/tech-policy/2012/09/googles-brazil-chief-detained-by-federal-police-over-youtube-video/.

189. Bishop, *supra* note 182.

190. *Complying with Brazil's Consumer Protection Code*, DIAZREUS (Jan. 27, 2012), http://diazreus.com/complying-with-brazils-consumer-protection-code/. *See* Lei No. 8.078, de 11 de Setembro de 1990, CÓDIGO CIVIL [C.C.] (Braz.).

191. Nicolo Zingales, *The Brazilian Approach to Internet Intermediary Liability: Blueprint for a Global Regime?*, 4 INTERNET POL'Y REV. 1 (2015).

192. *Id.*

193. Delfi AS v. Estonia, App. No. 64569/09, 2015 Eur. Ct. H.R. 69, https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-155105%22]}.

194. Although Facebook, Twitter and Instagram are dominate the Indian social media market, the autochthonous social network Hike is very popular among Indian Internet users. *See* Anish Gawande, *This App is Changing the Way Millions of Indians Use the Internet*, CNN (Aug. 3, 2017, 11:28 AM),

195. The Information Technology (Amendment) Act, 2008, No. 10 § 79, Acts of Parliament, 2009 (India).

196. Information Technology (Intermediaries Guidelines) Rules, 2011, Gazette of India, pt. II sec. 3(i), 13-14 (Apr. 11, 2011).

197. *Id.* § 3(2)(b).

According to the 2011 guidelines, intermediaries must follow "due diligence" to remove unlawful material.[198] The doctrine of due diligence comes from the realm of Indian business law and has many meanings, none of which have been clearly defined by a court or codified by a statute.[199] However, for the purposes of understanding the Indian philosophy behind intermediary liability, the doctrine of due diligence essentially means that once a company becomes aware that it is profiting off of the unlawful practices of a business partner, it must cease those unlawful business activities.[200] To illustrate this principle, in 2008, the Delhi High Court *in dicta* condemned a website that placed the maximization of profits over "[s]afeguard[ing] . . . prevailing moral values" in regard to its business model of profiting off of spreading links to obscene material.[201] Similarly, in July 2018, the Indian government warned the message-sharing platform WhatsApp that it could not escape accountability and responsibility for false rumors spread by users that led to the lynching deaths of 18 people.[202] Thus, the philosophy behind the Indian model of intermediary liability is similar to that of both the Brazilian model and the *Delfi* decision, whereby dutifully treating consumers is encouraged.[203]

H. Japanese Model

In Japan,[204] Act No. 137 of 2001 (the Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Senders, hereinafter referred to as the "Act") governs intermediary liability in Japan.[205] Article 3 of the Act explains that digital intermediaries "shall not be liable for any loss incurred" from an infringement of a user's rights (namely, to reputation and privacy), "unless where it is technically possible to take measures for preventing such information from being transmitted to unspecified persons."[206]

The Act is similar to § 230 in that it assures intermediaries that act to stop an infringement, whether proactively or upon receiving notice, do not open themselves up to liability.[207] However, like some other legal regimes discussed here, the Act

198. *Id.* § 3.

199. *See Legal Due Diligence*, INT'L FIN. L. REV. (July 12, 2001), http://www.iflr.com/Article/2027418/Legal-due-diligence.html.

200. *See, e.g.*, James Grandolfo, Ajit Sharma, Vandana Shroff, H. Jayesh & Kavita Mohan, *India*, 44 INT'L LAW. 663, 673 (2010).

201. Avnish Bajaj v. State, (2008) 150 DLT 279 (India).

202. *India Lynchings: WhatsApp Sets New Rules After Mob Killings*, BBC (July 20, 2018), https://www.bbc.com/news/world-asia-india-44897714.

203. *See* Rajinder Kaur & Rashmi Aggarwal, *Cyber Crime in India: An Analysis of the Regulatory Framework*, 20 COMPUT. & TELECOMM. L. REV. 17 (2014).

204. U.S. based Instagram, Twitter, and Facebook dominate the social media market in Japan. *See* Caylon Neely, *Japan's Top Social Media Networks for 2018*, HUMBLE BUNNY (Jan. 29, 2018), http://www.humblebunny.com/japans-top-social-media-networks-2018/. However, the autochthonous social messaging app Line is very popular in the country. *See* Jon Russell, *Understanding Line, the Chat App Behind 2016's Largest Tech IPO*, TECHCRUNCH, https://techcrunch.com/2016/07/14/understanding-line-the-chat-app-behind-2016s-largest-tech-ipo/ (last visited Oct. 7, 2018).

205. Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Senders Act, Law No. 137 of 2001 (Japan), http://www.unesco.org/culture/pdf/anti-piracy/Japan/Jp_%20LimitLiability_Telecom_en.

206. *Id.* art. 3(1).

207. *Id.* art. 3(2).

stipulates that intermediaries lose immunity from liability if they have knowledge of an infringement and do not take action.[208] Act No. 137 also goes a step further in empowering users by granting alleged victims of infringement the right to demand that intermediaries hand over information about the alleged infringing parties, including their names and addresses, provided that they have sufficient evidence to show the specific party infringed upon their rights.[209] In the spirit of due process, the intermediaries are further required to notify the alleged infringers that their information is being sought.[210] Thus, the Japanese approach to intermediary liability appears to balance competing interests of platforms and users in a spirit of social responsibility.

I. South Korean Model

In South Korea,[211] intermediary liability is rooted in Article 44 of the Act on the Promotion of Information and Communications Network Utilization and Information Protection, which gives a rather detailed account of what such liability must look like.[212] First, Article 44 broadly stipulates that "[n]o user may circulate any information violating another person's rights, including invasion of privacy and defamation, through an information and communications network," and "[e]very provider of information and communications services shall make efforts to prevent any [such] information . . . from being circulated through the information and communications network operated and managed by it."[213] Doing so will lead to the intermediary's liability being "mitigated or discharged."[214]

Second, Article 44 states that the victim of an "invasion of privacy or defamation" has the right to demand a deletion of the infringing content, as well as a right to a rebuttal, provided he or she can furnish evidence of the violation.[215] When such a situation arises, the intermediary must post a public notice on its platform that the situation occurred and that it is attempting to rectify it.[216] If the intermediary is unsure of whether the content was infringing, it can temporarily deny access to the content for up to 30 days.[217]

Third, Article 44 states that if an intermediary voluntarily removes content that is defamatory or invades a user's privacy, it will not have assumed liability for the content; this puts the Korean intermediary liability regime on par with several others

208. *Id.* art. 3(1)(ii).
209. *Id.* art. 4(1).
210. *Id.* art. 4(2).
211. Although Facebook and Twitter are gaining ground in the South Korean social media market, the autochthonous social networks KakaoStory and Cyworld retain very high levels of penetration among Korean Internet users. *See Explained: The Unique Case of Korean Social Media*, LINKFLUENCE (July 28, 2017), https://linkfluence.com/the-unique-case-of-korean-social-media/.
212. Act on Promotion of Information and Communications Network Utilization and Information Protection, Etc., Act No. 9119, June 13, 2008, art. 44 (S. Kor.), *translated in* Korea Legislation Research Institute online database, http://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=38422&type=sogan&key=41.
213. *Id.*
214. *Id.* art. 44-2(6).
215. *Id.* art. 44-2(1) (Request for Deletion of Information).
216. *Id.* art. 44-2(2).
217. *Id.* art. 44-2(4).

discussed here in terms of offering a "Good Samaritan" clause for platforms.[218] Finally, Article 44 stipulates that upon an order from the government, intermediaries must remove several types of unlawful content, including obscene material, defamatory messages, content that "arouses fear or apprehension," content that "divulges a state secret," and content that that "aides or abets in the committing of a criminal act."[219]

Until 2010, intermediaries were subject to liability if they failed to maintain procedures for unmasking users who fraudulently use another user's identity.[220] If intermediaries with "more than 100,000 users" did not routinely furnish this information when requested, they were mandated by a 2005 law to require users to post comments under their real names.[221] However, in 2012, the Korean Constitutional Court held that this provision of Article 44 was unconstitutional because it chilled individuals' anonymous speech without sufficiently fulfilling the provision's goal of minimizing harms resulting from anonymous speech.[222] Despite this ruling, the Korean model, like the Japanese model, mandates that intermediaries must "furnish the name and address" of a user who "defames or violates the privacy" rights of another user if the latter can prove he or she can prevail in a civil or criminal defamation proceeding.[223] The victim receiving the "information may not use the information for any purpose other than filing a civil or criminal complaint."[224]

In 2012, the Korean Constitutional Court upheld the constitutionality of Article 44, holding that the notice-and-takedown regime it established did not unreasonably infringe upon freedom of expression.[225] The court held Article 21 of the Korean Constitution,[226] which mandates that freedom of expression enjoyed by media companies must not infringe upon the individual rights of citizens (e.g., to honor and privacy) applied to digital intermediaries, and that the provisions of Article 44 imposed only minimal restrictions on freedom of expression.[227] The court also noted

---

218. *Id.* art. 44-3(1) (Discretionary Temporary Measures).
219. *Id.* art. 44-7(1) (Prohibition on Circulation of Unlawful Information).
220. *Id.* art. 44-5(1) (Verification of Identity of Users of Open Message Boards).
221. Public Official Election Act, Act No, 7681, Aug. 4, 2005 (S. Kor.), *translated in* Korea Legislation Research Institute online database, http://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=38422&type=sogan&key=41. *See* K.S. Park & Da Young Chung, *Mandatory Identity Verification in the Internet: Did Google Do the Right Thing?*, 5 KOR. U. L. REV. 203, 207 (2009).
222. Identity Verification System on Internet [Const. Ct.], 2010Hun-Ma252 (consol.), Aug. 23, 2012, (2012 KCCR, 590) (S. Kor.).
223. Act of Promotion of Information and Communications Network Utilization and Information Protection, Etc., Act No. 9119, Jun. 13, 2008, art. 44-6(1) (S. Kor.) (Claim to Furnish User's Information).
224. *Id.* art. 44-6(3).
225. Article 44-2. Paragraph 2 of the Act of Promotion of Information Network Usage and Information Protection, etc. [Const. Ct.], 2010Hun-Ma88 (consol.), May 31, 2012, (KCCR, 578) (S. Kor.).
226. Amended 1987 Daehanminkuk Hunbeob [HUNBEOB] [CONSTITUTION] art. 21(4) (S. Kor.):
(1) All citizens shall enjoy freedom of speech and the press, and freedom of assembly and association.
(2) Licensing or censorship of speech and the press, and licensing of assembly and association shall not be permitted.
(3) The standards of news service and broadcast facilities and matters necessary to ensure the functions of newspapers shall be determined by Act.
(4) Neither speech nor the press shall violate the honor or rights of other persons nor undermine public morals or social ethics. Should speech or the press violate the honor or rights of other persons, claims may be made for the damage resulting therefrom.
227. Article 44-2. Paragraph 2 of the Act of Promotion of Information Network Usage and Information Protection, etc. [Const. Ct.], 2010Hun-Ma88 (consol.), May 31, 2012, (KCCR, 578) (S. Kor.).

that Article 44 was good policy because it allowed digital intermediaries the opportunity to avoid costlier damages should they be found liable of facilitating UGC that either defamed or invaded the privacy of other users.[228] The court also stressed the purpose of Article 44 was to protect individual users who found themselves victims of infringing content, and that it could not be invoked to squelch speech critical of the government.[229]

In 2014,[230] and again in 2015, the Korean Constitutional Court also upheld the constitutionality of the provision of Article 44 requiring intermediaries to remove content that divulges state secrets when notified by the Korea Communications Commission.[231] All told, the South Korean approach to intermediary liability—much like the Japanese model—appears to balance the competing interests of consumer protection and due process for freedom of expression, and it affords platforms a reasonable degree of immunity from liability.

J. Concluding Thoughts on the Comparative Analysis

The nine models above are not a complete representation of how democratic polities conceive of the laws of intermediary liability in the context of extreme speech. However, they do reveal points along a spectrum. At one end, intermediaries are afforded exceptional immunity from liability; at the other, these platforms are subject to a strict notice-and-takedown regime when it comes to extreme UGC. Understanding this spectrum is key to distilling a set of principles for ethical operations by social media platforms.

The United States model of § 230 sits on the former end of the spectrum. Platforms in the United States only face liability for harmful UGC when they materially contribute to its creation.[232] This model affords platforms a high degree of control over users' content with little responsibility for it,[233] and some have argued this is the main impetus for the success of United States social networks and other platforms that traffic in UGC.[234] Moving along the spectrum, the UK and South African models share the goals of the United States model of promoting self-regulation among platforms, though the latter two models more actively encourage self-regulation through government-supported, non-binding codes of conduct.

On the opposite end of the spectrum are the Brazilian and Indian models, which see social networks as directly responsible for facilitating harmful UGC rather than mere neutral platforms. The Japanese and South Korean models fall somewhere in the middle of the spectrum, seeking to balance the goals of ensuring the commercial

228. *Id.* at 11.
229. *Id.* at 12.
230. Removal of Posts Containing Unlawful Information Case [Const. Ct.], 2012Hun-Ba325, Sept. 25, 2014, (2014 KCCR, 466) (S. Kor.).
231. Jinbo Network Center v. Korea Communications Commission [S. Ct.], 2012Du26432, Mar. 26, 2015 (S. Kor.).
232. Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008); Jones v. Dirty World Entm't Recordings, LLC, 755 F.3d 398 (6th Cir. 2014).
233. *See* Sandra Braman & Stephanie Lynch, *Advantage ISP: Terms of Service as Media Law*, 5 NEW MEDIA & SOC'Y 422 (2003).
234. *See, e.g.*, Ardia, *supra* note 109; Jonathan W. Peters & Brett G. Johnson, *Conceptualizing Private Governance in a Networked Society*, 18 NORTH CAROLINA J. L. & TECH. 15, 22–23 (2016); James Grimmelmann, *The Internet Is a Semicommons*, 78 FORDHAM L. REV. 2799 (2010). *See generally* LAWRENCE LESSIG, CODE: VERSION 2.0 (2006).

success of platforms, promoting robust participation by users, and protecting users from undue harm caused by UGC. Meanwhile, the EU model appears to be shifting more toward the Brazilian and Indian models by placing more liability on platforms out of a spirit of consumer protection. The German "Network Enforcement Act," as well as the 2015 *Delfi* decision, show the EU moving in this direction. Meanwhile, the messiness of the Australian model reveals how much in flux the issue of intermediary liability for harmful UGC is. Indeed, even the United States model is not immune from this trend toward skepticism of social media platforms.[235] Whether this trend results in the revision of § 230 remains to be seen. In the meantime, it would be wise for platforms to follow several ethical principles that can be distilled from the intermediary laws reviewed above. This article now turns to address those principles.

IV. SYNTHESIS: PLATFORM ETHICS INFORMED BY LAW

As I noted earlier, I have called elsewhere for scholars and industry leaders to conceptualize possible versions of "platform ethics," or the notion that digital intermediaries maintain some kind of duty to their users.[236] In particular, I argue that three factors should be taken into account when defining platforms ethics: (1) the extent to which intermediaries maximize the speaking power of individual users; (2) the extent to which intermediaries mitigate against unnecessary harm against users; and (3) the extent to which intermediaries facilitate the goals of democracy. Ultimately, the goal is for platforms and users to decide what is the proper balance. To this calculus, digital intermediaries can add the goals of avoiding legal regulation and streamlining costs.

This goal assumes that digital intermediaries are *media* institutions owing the same duty to democracy as newspapers and broadcasters, even though they might want to consider themselves merely tech companies that are value-neutral.[237] This proposition is not new. Before social media platforms were common phenomena, Professor Stephanie Craft called on large media conglomerates to recognize a duty to their audiences (i.e., the public) due in no small part to the notion that the framers of the Constitution "thought of the press as an entity whose purpose was not solely or even predominantly profit generation, but public service."[238] Similarly, Professors David Allen and Elizabeth Hindman have argued that media ethics should be a concept built on an institutional level, with the key focus of inquiry being how

---

235. *See, e.g.*, Dan Levine & Kristina Cooke, *Tech Industry's Legal Shield is Feeling the Heat*, REUTERS (Aug. 18, 2016, 12:05 AM), https://www.reuters.com/article/us-tech-court-idUSKCN10T0ET; Alfred Ng, *Tech Giants to Congress: Sorry About Our mistakes, But There's No Bias*, CNET (July 18, 2018, 7:18 AM), https://www.cnet.com/news/tech-giants-tell-congress-theyre-not-censoring-with-a-political-bias/; Nicholas Conlon, *Freedom to Filter Versus User Control: Limiting the Scope of § 230(c)(2) Immunity*, 2014 U. ILLINOIS J. L. TECH. & POL'Y 105, 109 (2014); CITRON, *supra* note 91.

236. *Speech, Harm and the Duties of Digital Intermediaries*, *supra* note 5, at 20.

237. *See* Brett G. Johnson & Kim Kelling, *Placing Facebook: "Trending," "Napalm Girl," "Fake News" and Journalistic Boundary Work*, 12 JOURNALISM PRACTICE 817 (2018) (finding that journalists have sought to discursively construct Facebook as a media institution rather than a tech company so that the company can be forced to abide by the same ethical principles as journalists when dealing with harmful third-party content).

238. G. Stuart Adam, Stephanie Craft & Elliot D. Cohen, *Three Essays on Journalism and Virtue*, 19 J. MASS MEDIA ETHICS 247, 265 (2004).

media companies devote their resources to the betterment of society and the fulfillment of democratic ideals.[239]

Professor Michael Perkins reminds us that the fundamental difference between law and ethics is that the "law sets a minimum standard below which our actions must not fall," while "ethics sets a higher standard to which we ought to aspire."[240] The legal regimes analyzed here offer some clues as to how platform ethics might be conceptualized through distilling moral values from legal principles that balance well with one another and lead intermediaries to engage in practices that go beyond minimum standards set by law.

First, platforms should recognize that they are dependent on users, not the other way around. In contrast to § 230, intermediary liability regimes in other parts of the world tend to impose less immunity to digital intermediaries. Namely, a "social theory of responsibility" in which "control capability [over content] implies co-responsibility" (morally, if not legally, speaking) distinguishes these regions from the United States.[241] The non-United States regimes view individual citizens as the most important stakeholders in a networked economy that thrives on facilitating public discourse, but from a different perspective than United States law.

Generally, these legal regimes tend to view digital intermediaries as being dependent upon individuals, and they acknowledge the fact that digital intermediaries commoditize the speech of individuals for profit.[242] For example, notions particularly present in the European, Brazilian, and Indian regimes of intermediary liability—that platforms have unscrupulous profit motives behind allowing third-party content, even some of the most harmful types—should translate into an ethical principle that the ability of intermediaries to profit off of individuals ends when that profit is steeped in harmful content. Therefore, notions of a duty of care, present in all of the regulatory models examined here except § 230, can translate into an ethical duty that intermediaries should follow in their relations with users.

Second, intermediaries should recognize the potential of users' speech to foster deliberative democracy online, and thus they should follow ethical precepts that would seek to turn their platforms into spaces where robust public debate can occur amid a genuine ethos of trust between users and the intermediaries. Such a position would transcend the "aggregational" concept of freedom of expression that platforms currently practice, a concept that fits well with the business model of intermediaries but not with their role in democracy.[243]

This second goal stems from the values enshrined particularly in the United States' § 230. Section 230 provides much greater protection for speech than the others analyzed here.[244] This easy distinction comes from the fact that the foreign approaches analyzed involve notice-based liability, whereas notice does not trigger

239. David S. Allen & Elizabeth Blanks Hindman, *The Media and Democracy: Using Democratic Theory in Journalism Ethics*, *in* THE ETHICS OF JOURNALISM: INDIVIDUAL, INSTITUTIONAL AND CULTURAL INFLUENCES 185, 186–203 (Wendy N. Wyatt ed., 2014).

240. Perkins, *supra* note 12, at 195.

241. Tomas A. Lipinski, Elizabeth A. Buchanan & Johannes J. Britz, *Sticks and Stones and Words that Harm: Liability vs. Responsibility, Section 230 and Defamatory Speech in Cyberspace*, 4 ETHICS & INFO. TECH. 143, 156 (2002).

242. *See* Ganaele Langlois, *Participatory Culture and the New Governance of Communication: The Paradox of Participatory Media*, 14 TELEVISION & NEW MEDIA 91 (2012).

243. *Facebook's Free Speech Balancing Act*, *supra* note 7, at 36.

244. *See* Pinto et al., *supra* note 120, at 5.

liability under § 230.[245] On the other hand, the somewhat paradoxical tradeoff of prioritizing the protection of speech on social media platforms is that the law may be affording so much protection to these intermediaries that they control more speech than § 230 intended.

Professor Sandra Braman argues that digital intermediaries, in hypocritical and self-contradictory fashion, "do not want to be content providers but do want to control all content," and that § 230 effectively gives them "*control without liability*."[246] Intermediaries seek to be "rewarded for facilitating expression but not liable for its excesses."[247] Similarly, Professor Rebecca Tushnet and Professor Dawn Nunziato argue that § 230 gives digital intermediaries too much of an incentive to control speech.[248] The argument goes that if intermediaries are able, under § 230, to proactively remove objectionable content without fear of liability, they will do so to the detriment of individuals' ability to speak freely on these platforms.[249] Tushnet accuses this legal regime of "simultaneously supporting freedom and suppression,"[250] and posits that "if we limit intermediary responsibility . . . we should also limit intermediary power to control speech."[251] In other words, digital intermediaries face a dilemma under the § 230 regime: they can protect speech and be accused of not doing enough to prevent harm, or they can remove harmful content at the request of individuals and be accused of not doing enough to protect speech.[252]

Third, platforms should act with transparency regardless of what decisions they make on governing third-party content. This goal is particularly evident in the South African ISPA Code, but it is also present in the Japanese and South Korean provisions requiring intermediaries to notify users who posted infringing content that their content was being removed or that their identities were being revealed to the alleged victims of the content.[253] Even in the rather draconian German Network Enforcement Law, the spirit of transparency behind filing records with a government agency on how platforms handle unlawful third-party content should live on in a conception of platform ethics.

What would these ethical principles look like in practice? For starters, although it might seem that following an ethical policy of mitigating harmful third-party content would require regular monitoring, I do not propose that platforms adopt the

245. *See supra* notes 103–110.

246. Braman & Lynch, *supra* note 233, at 438 (emphasis in original).

247. Tarleton Gillespie, *The Politics of 'Platforms'*, 12 NEW MEDIA & SOC'Y 347, 356 (2010).

248. Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986, 1003 (2008); DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE 36 (2009).

249. NUNZIATO, *supra* note 248, at 2–3.

250. Tushnet, *supra* note 248, at 1010–11.

251. *Id.* at 1009.

252. *Id.*; NUNZIATO, *supra* note 248.

253. *See Code of Conduct*, INTERNET SERVICE PROVIDERS' ASS'N J(31) (June 1, 2016) (S. Afr.), https://ispa.org.za/code-of-conduct/; Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Senders Act, Law No. 137 of 2001, art. 4, para. 2 (Japan), http://www.unesco.org/culture/pdf/anti-piracy/Japan/Jp_%20LimitLiability_Telecom_en; Act on Promotion of Information and Communications Network Utilization and Information Protection, Etc., Act No. 9119, June 13, 2008, art. 44-6(1) (Claim to Furnish User's Information) (S. Kor.), *translated in* Korea Legislation Research Institute online database, http://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=38422&type=sogan&key=41.

ethical version of the monitoring system that the ECtHR called for in *Delfi v. Estonia*.[254] Rather, a notice-and-takedown regime whereby users submit requests to platforms for removal of harmful content would meet this standard. The practice that would take platforms to a higher level of moral action beyond minimum legal requirements for removing content when notified would be for intermediaries to *promote* civil and constructive discourse on matters of public concern rather than merely *delete* harmful speech.

For instance, if a Twitter user tweets something that is racist, and another user flags the tweet for takedown, Twitter, if it chooses to remove the tweet, could respond to the user in the following manner:

> We have decided to remove your tweet because several other users found its content racist and offensive. Our goal at Twitter is to promote a civil and constructive discourse among users. We value your contributions to this discourse, and we encourage you to keep tweeting in a way that respects the values of our diverse community. If you believe the tweet was removed in error, we invite you respond to the removal and let us know why you believe your tweet was valuable to public discourse.

The next step in the process would be for intermediaries to act transparently and pull back the curtain to reveal to users what the process of handling user requests for removal of content looks like. This approach could be based on several elements that are steeped in an ethic of due process. Platforms should show users the same guides that the employees rely on for taking down content.[255] Indeed, Facebook took this step in April 2018 by releasing to the public the guidelines that their content moderators use.[256] Platforms should allow the user whose content was removed to see, at the very least, how many times their content was flagged. Out of concern for the safety for the individuals doing the flagging, it would be best to not allow the user to see who flagged his or her content, lest he or she retaliates in some way. However, knowing the number of flags would help the user get a sense of the severity of his or her content.

Platforms should also allow users the option to appeal to get the posts reinstated. This model would be similar to the Digital Millennium Copyright Act's provision that a user can formally appeal to have their allegedly copyright-infringing content be reinstated under the theory that the takedown request was issued in bad faith.[257] This would give the user the opportunity to critically reflect on his or her content to make the case that it should be a part of a social network's public discourse. Facebook also instituted an appeals process for removed content in April

254. Delfi AS v. Estonia, App. No. 64569/09, 2015 Eur. Ct. H.R. 60–61, https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-155105%22]}.

255. *See*, *e.g.*, Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads'*, GAWKER (Feb. 16, 2012, 3:45 PM), http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads (including an example of such a guide that an employee leaked to the press).

256. Monika Bickert, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, FACEBOOK NEWSROOM (Apr. 24, 2018), https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/.

257. 17 U.S.C. § 512(g) (2010).

2018,[258] though it has received criticism for its lack of transparency, which is ironic given that the social network had simultaneously revealed its content moderation guidelines to the public.[259]

Certainly, achieving a more complete ethic of transparency will take some time and growing pains. However, scholars are starting to build more sophisticated ideas of how to make content governance more transparent. For example, communication scholar Tarleton Gillespie has suggested that Facebook and other platforms institute a practice that he calls "collective flagging," whereby data created by users reacting to extreme content could be viewed in tandem with that content.[260] Gillespie lays the foundation for this idea in the notion that this data essentially belong to users, because they would create the flags.[261]

It is worth noting that the legal regimes that establish a duty of care on intermediaries when they are notified of harmful content go no further in their definitions of harm than Smolla's three categories. Societal-level harms, such as those that are potentially caused by fake news, are absent from these laws, although there is no reason to believe laws could not be amended to force intermediaries into action on such content. As of this writing, none of the laws analyzed here have been amended, likely due to the relative recentness of the phenomenon and the lack of certainty as to what degree (or even whether) it causes harm.[262] In the meantime, intermediaries could preempt legal action by setting up what I call a "notice-and-discussion" regime to handle UGC that is based on fake news. This regime is similar to the scenario discussed above in that it would be designed to encourage civil and constructive public discourse.

For instance, if Uncle John links to a fake news story while making a political diatribe on Facebook, Cousin Jane, who knows the story is fake, could notify Facebook of the presence of the story as a means to prompt Facebook to post a comment to Uncle John's post with evidence demonstrating that the story is fake. The goal here would not be to humiliate Uncle John, but rather to supply evidence from a "neutral" third party that the story is fake. Furthermore, Facebook's post could say something like the following:

> We appreciate your passion for this issue and your desire to start a conversation about it on Facebook. We encourage you to continue this discussion with your friends armed with the appropriate facts about the issue.

Facebook could then supply Uncle John, and anyone else following the discussion, with links to a wide array of news sites that discuss an issue related to the issue at the heart of the fake news story (e.g., an election). The goal of this approach would be to stanch the flow of fake news, while not simultaneously cutting off the

258. Bickert, *supra* note 256.

259. Kate Klonick, Opinion, *Facebook Released Its Content Moderation Rules. Now What?*, N.Y. TIMES (Apr. 26, 2018), https://www.nytimes.com/2018/04/26/opinion/facebook-content-moderation-rules.html?action=click&pgtype=Homepage&clickSource=story-heading&module=opinion-c-col-right-region&region=opinion-c-col-right-region&WT.nav=opinion-c-col-right-region.

260. TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 199 (2018).

261. *Id.*

262. Hunt Allcott & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*, 31 J. ECON. PERSP. 211, 232 (2017) (finding little evidence that fake news played a major role in determining the results of the 2016 U.S. presidential election).

discussion of important matters of public concern. In other words, the goal is more speech based on facts.

It should be readily apparent that social media platforms want to embrace a system of platform ethics, at the very least for prudential reasons. Professor Monroe Price and researcher Stefaan Verhulst remind us that ethical principles can help digital intermediaries "evolve according to changing norms" in our global society.[263] According to Price and Verhulst, ethical principles "neither mirror public opinion nor necessarily lag or lead public opinion in terms of cultural norms"; rather, they "represent a temporarily agreed upon set of standards which serves as a way-station or *modus vivendi* as new modes of information are distributed."[264] The authors note that "[g]reater flexibility can make it easier to respond to changes in technology [and] modify expectations and outcomes," which is especially important in the context of harmful third-party content, "which is culturally diverse and subject to changing norms—is better suited to self-regulation."[265]

Furthermore, Price and Verhulst argue that self-regulation via the adoption of ethical standards affords intermediaries the "benefit of avoiding state intervention in sensitive areas of basic rights."[266] Furthermore, a constant concern of digital intermediaries is that they might lose their "coolness" status among their users, leading users to turn their eyeballs elsewhere.[267] The adage that a platform's competition "is just a click away" may be clichéd,[268] but its truth lies in the once-popular social networking platforms that have since been shuttered, such as MySpace, Friendster, and Orkut. Following a set of precepts from platform ethics could help intermediaries at least maintain, or even enhance, their coolness as they seek to engage more concertedly and transparently with their users and with their vital role in democracy.

V. CONCLUSION

This article examined intermediary liability laws in nine liberal democratic polities as they pertain to extreme UGC. The findings from this review were applied to further build upon the concept of "platform ethics."[269] The debate over what duties digital intermediaries have to their users is expanding and maturing.[270] It must be approached using metaethical theories as well as ethical norms distilled from comparative legal analysis. Platform ethics also must be conceptualized vis-à-vis the various ways in which platforms affect our lives. Extreme speech is only part of the story. Data privacy, advertising, and the facilitation of journalism are also important

263. MONROE E. PRICE AND STEFAAN G. VERHULST, SELF-REGULATION AND THE INTERNET 43 (2005).
264. *Id.*
265. *Id.* at 35.
266. *Id.* at 9.
267. Ally Marotti, *Younger Users Flee their Parents' Favorite Social Network, Facebook, at Surprising Pace*, CHICAGO TRIB. (Feb. 12, 2018, 11:40 AM), http://www.chicagotribune.com/business/ct-biz-younger-users-leaving-facebook-20180212-story.html.
268. *See* Steve Lohr, *Onetime Allies in Antitrust Part Ways over Google*, N.Y. TIMES (Dec. 16, 2012), https://www.nytimes.com/2012/12/17/technology/onetime-allies-in-antitrust-part-ways-over-google.html (attributing this phrase to Google, which "repeatedly says, competition is "'just a click away.'").
269. *Speech, Harm and the Duties of Digital Intermediaries*, *supra* note 5.
270. *See* Johnson & Kelling, *supra* note 237 (finding that journalists have ascribed to Facebook the same moral duties to democratic society that generally apply to the press).

lenses from which to view platform ethics. Scholars should continue the conversation in these areas through scholarship that involves legal research as well as ethical research, empirical studies as well as normative ones, and adopting domestic approaches to study as well as comparative.

As a final note, the discussion on whether the exceptional § 230 should be revised is also still ongoing. This article does not expressly advocate for a position in that debate, except to say any amendments that do materialize should not be put forth as a kneejerk reaction to the harms social media platforms facilitate against both individuals and society. Rather, they should be considered with a full understanding of the nature of the harms being facilitated and a high degree of certainty that amendments would in fact curb these harms. If platforms have not developed a set of sound ethical principles (not unlike those discussed here) to ameliorate these harms on their own, then United States lawmakers would be justified in amending § 230 to do so.