EVALUATING THE APPROACHES TO SOCIAL MEDIA LIABILITY FOR PROHIBITED SPEECH

ZI EN CHOW

I. INTRODUCTION

> "Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet . . . . The extent to which Facebook posts and messages have led to real-world discrimination and violence must be independently and thoroughly examined."[1]

In recent years, Internet intermediary responsibility for prohibited content posted by the users of those intermediaries has been in the spotlight. U.N. human rights experts have suggested that Facebook played a role in spreading hate speech in Myanmar, contributing to large scale human rights violations.[2] Ultra-nationalist Buddhists in Myanmar have utilized Facebook to incite violence against the Rohingya people and

1. Human Rights Council, Rep. of the Indep. Int'l Fact-Finding Mission on Myan., ¶ 74, U.N. Doc. A/HRC/39/64 (Sept. 12, 2018).

2. Tom Miles, *U.N. Investigators Cite Facebook Role in Myanmar* Crisis, REUTERS, Mar. 12, 2018, https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN; *id.*; *see* Louise Matsakis, *Twitter Releases New Policy on 'Dehumanizing Speech,'* WIRED (Sept. 25, 2019, 9:00 AM), https://www.wired.com/story/twitter-dehumanizing-speech-policy (noting that Facebook has been accused of facilitating the Myanmar genocide).

other ethnic minorities.[3] In an April 2018 hearing, the U.S. Congress questioned Mark Zuckerberg, Chairman and Chief Executive Officer of Facebook, about Facebook's complicity in the violence in Myanmar. In response, Zuckerberg assured Congress that Facebook was working to hire more Burmese-language content reviewers to better tackle hate speech in Myanmar.[4] Thousands of kilometers away, the question of intermediary responsibility for user-generated content also arose in Kenya as Facebook was filled with ethnically fueled hate speech and propaganda during the 2017 Kenyan election, during which about a hundred Kenyans lost their lives in election-related violence.[5] In South Sudan, the civil war has killed tens of thousands and created two and a half million refugees as of 2018, with Facebook exacerbating the conflict in hosting hateful speech on its platform.[6] The general perception, as echoed by multiple agencies, is that Facebook and other social media giants are unprepared and ill-equipped to regulate user-generated content on their platforms.[7]

---

3. Miles, *supra* note 2.

4. David Z. Morris, *There's No Easy Tech Fix for Online Hate Speech*, SLATE (Apr. 19, 2018, 9:41 AM), https://slate.com/technology/2018/04/facebook-shouldnt-count-on-artificial-intelligence-to-fix-its-hate-speech-problems.html.

5. Abigail Higgins, *Facebook Doesn't Need to Engineer World Peace, But It Doesn't Have to Fuel Violence*, BRIGHT MAG. (Apr. 6, 2018), https://brightthemag.com/facebook-cambridge-analytica-south-sudan-myanmar-data-trump-kenya-violence-hate-speech-fake-news-61eb39e425bf; Drazen Jorgic, *Kenya Tracks Facebook, Twitter for Election "Hate Speech,"* REUTERS, Feb. 5, 2013, https://www.reuters.com/article/net-us-kenya-elections-socialmedia/kenya-tracks-facebook-twitter-for-election-hate-speech-idUSBRE9140IS20130205.

6. Higgins, *supra* note 5. **R**

7. *See* David Goldman, *Big Tech Made the Social Media Mess. It Has to Fix It*, CNN (Oct. 29, 2018), https://www.cnn.com/2018/10/29/tech/social-media-hate-speech/index.html (calling on social media companies to make efforts to limit the damage they cause); Sheera Frenkel, Mike Isaac & Kate Conger, *On Instagram, 11,696 Examples of How Hate Thrives on Social Media*, N.Y. TIMES (Oct. 29, 2018), https://www.nytimes.com/2018/10/29/technology/hate-on-social-media.html (noting that recent events have demonstrated that social media platforms are unable to handle disinformation and hate speech); Sam Levin, *Google to Hire Thousands of Moderators After Outcry Over YouTube Abuse Videos*, GUARDIAN (Dec. 5, 2017), https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos (reporting the Google will hire new moderators after its machine technology failed to detect and remove YouTube child abuse videos).

In light of the above, what kind of liability should governments impose on social media companies for hosting prohibited content on their platforms? There are three possible approaches: to not hold them liable at all; to hold them strictly liable in all instances; or to hold them liable in only certain instances.

This paper first examines the definition of prohibited speech according to regional and international instruments, before exploring how social media platforms currently regulate prohibited speech. It then analyzes the aforementioned three approaches that governments can take when dealing with social media and prohibited speech. The key argument is that social media platforms should only be held liable for failing to take down prohibited speech after receiving a court order to do so, as it is the most accountable, legitimate, and sustainable solution. Further, this paper proposes that adopting a partnership model between governments and social media platforms would effectively supplement the limited liability model.

II. Definition of Prohibited Speech

This paper will use the definition of prohibited speech from the International Covenant on Civil and Political Rights (ICCPR), which is the most widely recognized international treaty on human rights.[8] According to the ICCPR, prohibited speech refers to speech that governments can legitimately restrict on certain grounds. While the ICCPR protects an individual's freedom of expression, stating explicitly that "everyone shall have the right to freedom of expression,"[9] the ICCPR also recognizes that such freedoms can be restricted on the basis of the "rights or reputations of others," and for "the pro-

8. United Nations International Covenant on Civil and Political Rights art. 19(2), Mar. 23, 1976, 999 U.N.T.S. 171 ("Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.") [hereinafter ICCPR]. The Covenant has received widespread acceptance with 172 state parties and 6 signatories. *Status of Ratification: International Covenant on Civil and Political Rights*, U.N. Office of the High Comm'r for Hum. Rts., http://indicators.ohchr.org (last updated Apr. 15, 2019).

9. ICCPR, *supra* note 8.

tection of national security or of public order . . . , or of public health or morals."[10] Further, the ICCPR provides that state governments shall prohibit by law any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence.[11] The United Nations Human Rights Committee (UNHRC), which is the adjudicatory body for the ICCPR, has interpreted the above provisions to imply that governments can restrict speech if that restriction is prescribed by law, pursues one of the listed legitimate aims, and is necessary.[12]

Other regional instruments—such as the European Convention for the Protection of Human Rights and Fundamental Freedoms, the American Convention on Human Rights, and the African Charter on Human and Peoples' Rights—also impose similar restrictions on an individual's right to free expression.[13] The regional courts to those instruments have also held

---

10. *Id.* at art. 19(3)(a).

11. *Id.* at art. 19(3)(b).

12. For examples of this interpretation in practice, see Frank La Rue (Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression, Frank La Rue*, ¶ 29, U.N. Doc. A/HRC/23/40 (Apr. 17, 2013); Frank La Rue (Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression, Frank La Rue*, ¶ 24, U.N. Doc. A/HRC/17/27 (May 16, 2011) [hereinafter U.N. Doc A/HRC/17/27]; Frank La Rue (Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Opinion and Expression*, ¶ 15, U.N. Doc. A/66/290 (Aug. 10, 2011); Human Rights Comm., General Comment No. 34, ¶ 35, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011); Human Rights Comm., Communication No. 1022/2001, Velichkin v. Belarus, ¶ 7.3, U.N. Doc. CCPR/C/85/D/1022/2001 (Nov. 23, 2005); Human Rights Comm., Communication No. 736/1997, Malcolm Ross v. Canada, ¶ 11.2, U.N. Doc. CCPR/C/70/D/736/1997 (Oct. 26, 2000); Human Rights Comm., Communication No. 518/1992, Jong-Kyu Sohn v. Republic of Korea, ¶ 10.4, U.N. Doc. CCPR/C/54/D/518/1992 (Aug. 3, 1995); Human Rights Comm., Communication No. 458/1991, Womah Mukong v. Cameroon, ¶ 9.7, U.N. Doc. CCPR/C/51/D/458/1991 (Aug. 10, 1994).

13. Convention for the Protection of Human Rights and Fundamental Freedoms art. 10(2), *opened for signature* Nov. 4, 1950, E.T.S. No. 005. The relevant provision is as follows: "(2) The exercise of [the freedom of expression], since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national secur-

that governments must satisfy the three requirements that the UNHRC identified for a justifiable restriction on freedom of expression.[14]

---

ity, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary." American Convention on Human Rights "Pact of San Jose, Costa Rica" art. 13, Nov. 22, 1969, 1144 U.N.T.S. 123. The relevant provision of Article 13 is as follows: "(2) The exercise of the right [to freedom of thought and expression] shall not be subject to prior censorship but shall be subject to subsequent imposition of liability, which shall be expressly established by law to the extent necessary to ensure: (a) respect for the rights or reputations of others; or (b) the protection of national security, public order, or public health or morals." Article 13 additionally includes further limits beyond those listed in the ICCPR: "(4) Notwithstanding the provisions of paragraph 2 above, public entertainments may be subject by law to prior censorship for the sole purpose of regulating access to them for the moral protection of childhood and adolescence. (5) Any propaganda for war and any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin shall be considered as offenses punishable by law." African Charter on Human and Peoples' Rights art. 9(2), Jun. 27, 1981, 1520 U.N.T.S. 217. The relevant provision reads: "Every individual shall have the right to express and disseminate his opinions within the law."

14. For cases concerning the European Convention for the Protection of Human Rights and Fundamental Freedoms, see Perinçek v. Switzerland, App. No. 27510/08, HUDOC, ¶ 124 (2015), http://hudoc.echr.coe.int/eng?i=001-158235; Murat Vural v. Turkey, App. No. 9540/07, HUDOC, ¶ 1 (2015), https://hudoc.echr.coe.int/app/conversion/docx/pdf?library=ECHR&id=001-158235&filename=CASE%20OF%20PER%C4%B0N%C3%87EK%20v.%20SWITZERLAND.pdf&logEvent=false; Ceylan v. Turkey, App. No. 23556/94, 1999-IV Eur. Ct. H.R. 42, 42 (1999). For cases concerning the American Convention on Human Rights, see Inter-American Comm'n on Human Rights, Office of the Special Rapporteur for Freedom of Expression, Freedom of Expression and the Internet, OEA/Ser.L/V/II.CIDH/RELE/INF. 11/13, ¶¶ 58–64 (Dec. 31, 2013); Herrera-Ulloa v. Costa Rica, Preliminary Objections, Merits, Reparations and Costs, Judgment, Inter-Am. Ct. H.R, (ser. C), No. 107, ¶ 2 (July 2, 2004); Francisco Martorell v. Chile, Case 11.230, Inter-Am. Comm'n H.R., Report No. 11/96, OEA/Ser.L/V/II.95 doc. 7 rev. at ¶ 55 (1996). For cases concerning the African Charter on Human and Peoples' Rights, see African Comm'n on Human and Peoples' Rights, Zim. Lawyers for Human Rights & Institute for Human Rights & Dev. in Africa v. Zimbabwe, ¶ 80, A.H.R.L.R. 268 Comm. No. 294/04 (Apr. 3, 2009); African Comm'n on Human and Peoples' Rights, Interights v. Mauritania, A.H.R.L.R. 87 Comm. No. 242/2001, ¶¶ 78–79 (2004); African Comm'n on Human and Peoples' Rights Res. 62(XXXII)02, princ. II, Reso-

Out of the three requirements, the most challenging one for the courts to adjudicate is whether a restriction was necessary, given the fact-sensitive nature of such a determination.[15] In response, U.N. organs have identified some factors to introduce clarity and structure to the process.[16] However, there is no suggested guidance on how to weigh the factors relative to each other if they point to different conclusions.

III. Understanding Social Media Platforms

In the past fifteen years, social media platforms have radically transformed the media landscape and become closely identified with the freedom of expression. In the past, citizens had limited opportunities to express their views on media platforms. Traditional media outlets were typically controlled by the government or wealthy elites, and only the editors or producers wielded the power to express views through a mass medium. Today, anyone can create a social media account and share their opinions with the world, at no cost and with few restrictions on what they can post online. As such, social media has decentralized the means of communicating with the masses by creating a new virtual space for free expression with

---

lution on the Adoption of the Declaration of Principles of Freedom of Expression in Africa (Oct. 23, 2002).

15. With regard to the first requirement, an interference is generally prescribed by law if there is a sufficiently precise statute permitting the restriction of the individual's freedom of expression. With regard to the second requirement, interferences are also usually found to be in pursuit of a legitimate aim. U.N. High Comm'r for Human Rights, *Annual Rep. of the U.N. High Comm'r for Human Rights on the Expert Workshops on the Prohibition of Incitement to Nat'l, Racial or Religious Hatred*, ¶ 18, U.N. Doc. A/HRC/22/17/Add.4, app. (Jan. 11, 2013) [hereinafter Rabat Plan of Action].

16. The most prominent example is the Rabat Plan of Action, which list the factors for deciding if a specific speech act should be prohibited. Factors include (a) the intention of the speaker, (b) the content and form of the speech, (c) the context in which the speech was made, (d) the status of the speaker, (e) the likelihood of hatred, discrimination, or violence occurring, and (f) the extent of the speech act. *Id.* ¶ 29. Further, this framework was endorsed by the United Nations Office of the High Commission for Human Rights and the United Nations Human Rights Council. Farida Shaheed (Special Rapporteur in the Field of Cultural Rights), *Rep. of the Special Rapporteur in the Field of Cultural Rights, Farida Shaheed*, ¶ 89, UN Doc A/HRC/23/34 (Mar. 14, 2013); *UN Launches the Rabat Plan of Action*, Int'l Just. Resource Ctr. (Feb. 25, 2013), http://www.ijrcenter.org/2013/02/25/un-launches-the-rabat-plan-of-action.

a global audience. Thus, it comes as no surprise that access to social media is now an essential aspect of the public's understanding of free expression.

These developments pose tough questions on how online content should be regulated. Before anything is published in traditional media outlets, the content undergoes several rounds of editing and approval.[17] The editorial process ensures that the content is appropriate to be published, and serves as a bulwark against discriminatory or incendiary content. In contrast, social media platforms allow anyone to post content instantaneously on the Internet, with barely any approval process prior to publishing it. As such, any regulation of content is reactionary rather than preemptive. Facebook users must flag a post before human moderators examine the post and determine if the content is inappropriate.[18] For first-time violations, Facebook deletes the prohibited content and temporarily disables the account. For subsequent violations, Facebook disables posting rights for a longer period, or permanently suspends the account.[19] Another hurdle that social media platforms face in regulating content is the sheer volume of content that requires review. While specific numbers are not available, reports suggest that Facebook moderators review millions of user-flagged posts, groups, or pages every week.[20] Beyond just user-reported posts, moderators also have to sift through posts that Facebook's automated systems flag.[21] A former Facebook moderator disclosed that she reviewed an average of 8,000 posts a day, with less than 10 seconds to make a

17. Lauren McMenemy, *Run It Like A Newsroom: Turning Content Strategy into a Slick Operation*, SKYWORD (Aug. 29, 2017), https://www.skyword.com/contentstandard/marketing/run-like-newsroom-turning-content-strategy-slick-operation.

18. Dave Gershgorn, *Mark Zuckerberg Just Gave a Timeline for AI to Take Over Detecting Internet Hate Speech*, QUARTZ (Apr. 10, 2018), https://qz.com/1249273/facebook-ceo-mark-zuckerberg-says-ai-will-detect-hate-speech-in-5-10-years.

19. Harper Neidig, *Twitter Launches Hate Speech Crackdown*, HILL (Dec. 18, 2017), https://thehill.com/policy/technology/365424-twitter-to-begin-enforcing-new-hate-speech-rules.

20. Heather Kelly, *Facebook Reveals Its Internal Rules for Removing Controversial Posts*, CNN (Apr. 24, 2018, 5:57 AM), https://money.cnn.com/2018/04/24/technology/facebook-community-standards/index.html.

21. *Id.*

decision on whether the post should be removed.[22] As more people join social media platforms like Facebook or Twitter, this task of sifting through potentially inappropriate content will only get more difficult.

The consequence of a reactionary approach to content regulation is that prohibited content remains online for at least some time. Facebook only removed a video showing the murder of Robert Godwin Sr. in Ohio two hours after it was posted. Facebook did not catch the prior video in which the murderer shared his intention to murder, nor did any users report it.[23] Just a week later, Facebook removed a video of a man killing his daughter in Thailand only after the Thai Ministry of Digital Economy requested the removal a day later.[24] It is worth noting that social media platforms do not bear the cost of the presence of such material—the traumatized do.

Looking to the future, social media platforms are unlikely to become significantly faster in identifying and removing prohibited content. Facebook is developing artificial intelligence tools that can identify questionable speech and flag it to human moderators. However, Mark Zuckerberg himself admitted that such tools remain an imperfect solution. Determining whether a post constitutes prohibited speech is a linguistically nuanced undertaking,[25] one that artificial intelligence is poorly equipped to handle.[26] Although Facebook announced its plans to hire 10,000 more human moderators in 2018,[27] this number is minuscule compared to the ever-growing amount of content to screen.

---

22. Jo Ling Kent, Chiara Sottile & Alyssa Newcomb, *Monitoring Fake News Was Never a Priority, Says Ex-Facebook Worker*, NBC NEWS (Jan. 17, 2018), https://www.nbcnews.com/tech/social-media/monitoring-fake-news-was-never-priority-says-ex-facebook-worker-n838371.

23. Alyssa Newcomb, *Murdered Ohio Grandfather's Family Sues Facebook For Not Detecting Killer's Intent*, NBC NEWS (Jan. 31, 2018), https://www.nbcnews.com/tech/tech-news/murdered-ohio-grandfather-s-family-sues-facebook-not-detecting-killer-n843371.

24. Hannah Kuchler & Madhumita Murgia, *Facebook Removes Video of Thai Man Killing Baby Daughter*, FIN. TIMES (Apr. 25, 2017), https://www.ft.com/content/5c748ebe-2a1f-11e7-bc4b-5528796fe35c.

25. Morris, *supra* note 4. **R**

26. *See* Gershgorn, *supra* note 18 (describing why artificial intelligence sometimes struggles to detect hate speech).

27. Morris, *supra* note 4. **R**

The inadequacy of tools to regulate speech is not limited to Facebook. Twitter's process for detecting and removing prohibited speech remains unclear, except that it has a team reviewing content on its platform. Twitter reported in July 2017 that it employed a total of 3,200 staff, without clarifying how many are content moderators.[28]

A. *First Option: No Liability for Social Media Platforms*

The first option that governments have is to not hold social media platforms liable for prohibited speech on their platforms. The United States encourages social media platforms to block offensive content without holding them responsible for doing so ineffectively.[29] Internet intermediaries, which include social media platforms, are protected from liability for material that users post on their platforms because they are not treated as a "publisher" of the third-party content.[30] This paper previously highlighted the editorial process, or lack thereof, as a key distinction between traditional media and social media. Since social media platforms do not edit or approve content before users publish it online, they clearly cannot qualify as publishers in the way that the law intended.

---

28. Dave Gershgorn & Mike Murphy, *Facebook Is Hiring More People to Moderate Content than Twitter Has at Its Entire Company*, QUARTZ (Oct. 12, 2017), https://qz.com/1101455/facebook-fb-is-hiring-more-people-to-moderate-content-than-twitter-twtr-has-at-its-entire-company.

29. Communications Decency Act of 1996, 47 U.S.C. § 230(c) (2012); Issie Lapowsky, *Lawmakers Don't Grasp the Sacred Tech Law They Want to Gut*, WIRED (Jul. 17, 2018), https://www.wired.com/story/lawmakers-dont-grasp-section-230. A reproduction of the relevant provision of the Communications Decency Act is as follows:

"(c) Protection for "Good Samaritan" Blocking and Screening of Offensive Material Treatment of Publisher or Speaker: No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

Civil Liability: 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 or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or

any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1)."

30. Communications Decency Act, *supra* note 29. **R**

For instance, the U.S. Court of Appeals for the First Circuit held that a website hosting sex advertisements featuring children forced into prostitution was immune from liability because it was not considered a "publisher" of that third-party content.[31] In another case, the U.S. District Court for the Eastern District of New York held that Facebook was not liable even though Palestinian terrorists were using Facebook's platform to incite, enlist, organize, and dispatch individuals to target Israelis because Facebook was similarly not a "publisher" of that content.[32] Further, social media platforms are protected even if they attempt to moderate content, as civil liability cannot be imposed on them for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[33]

This approach ensures that individuals' freedom of expression receives the widest berth possible, since social media platforms are under no legal obligation to regulate prohibited content. However, this also implies that prohibited speech will stay online longer. Governments have to do the onerous legwork in uncovering these cases and requesting that the platforms take down content, especially when they are not even in an optimal position to detect potential cases. If social media companies today are struggling to stem the waves of prohibited content, it would be foolish to think that governments can do any better. Here, the cost of taking more time is that such prohibited content may have gone viral, thereby hurting people and potentially inciting violence against specific groups. Furthermore, social media companies benefit from the advertising dollars generated through their popularity amongst the masses, but inherit no risks from hosting their users' content. There is a compelling argument here that they should bear the costs arising from the risks that accompany user-generated content.[34]

---

31. Doe v. Backpage.com, LLC, 817 F.3d 12, 22 (1st Cir. 2016).

32. Cohen v. Facebook, Inc., 252 F. Supp. 3d 140, 160 (E.D.N.Y. 2017).

33. Communications Decency Act, *supra* note 29. **R**

34. Delfi AS v. Estonia, App. No. 40287/98, HUDOC, ¶¶ 112-113, 126 (2015), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-155105&filename=001-155105.pdf.

B. *Second Option: Strict Liability for Social Media Platforms*

The second option governments have is to hold social media companies strictly liable for all instances of prohibited content on their platforms. Under this paradigm, social media platforms are responsible because they facilitate the spread of hateful and incendiary content to a wider audience, thereby enabling the content to further promote imminent violence against certain groups. Without social media, such prohibited content would pose a far smaller risk of danger of unlawful action.

For example, in Thailand, social media platforms are held liable for intentionally supporting or consenting to prohibited speech on their platforms. If a platform failed to detect and remove such content in a timely manner, its consent will be implied.[35] In China, platforms that fail to monitor user activity, take down prohibited content, or report violations may face penalties such as fines, criminal liability, and revocation of business or media licenses.[36] Recent German legislation, Network Enforcement Act, requires platforms to remove illegal, racist, or slanderous posts within twenty-four hours of receiving a user notification; failure to do so could result in an astronomical fine of up to 50 million euros.[37]

While such an approach incentivizes social media companies to proactively ensure that prohibited content is removed, such an obligation may be too onerous for these companies and is likely to impinge on individuals' freedom of expression.

In the earlier section, this paper explained the difficulties social media companies face in rapidly identifying and take prohibited content down. Hiring more human moderators is not a sustainable solution, as the number of social media users

---

35. KHEMTHONG TONSAKULRUNGRUANG, ONLINE INTERMEDIARY LIAB. RESEARCH PROJECT AT UNIV. OF WASH. SCH. OF LAW, STUDY OF INTERMEDIARY LIABILITY IN THAILAND: HATE SPEECH 24 (2015).

36. REBECCA MACKINNON, ET AL., U.N. EDUC., SCI. & CULTURAL ORG., FOSTERING FREEDOM ONLINE: THE ROLE OF INTERNET INTERMEDIARIES 40 (2014).

37. Melissa Eddy & Mark Scott, *Delete Hate Speech or Pay Up, Germany Tells Social Media Companies*, N.Y. TIMES (June 30, 2017), https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html; Philip Oltermann, *Tough New German Law Puts Tech Firms and Free Speech in Spotlight*, THE GUARDIAN (Jan. 5, 2018), https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

grows exponentially every year. Moderators also experience trauma from having to view thousands of pieces of horrendous or vitriolic content every day, clouding their judgment and increasing turnover rates.[38] Finally, human moderators also struggle with borderlines cases, knowing they will face criticism for either choice they make.[39] For example, Twitter's permanent suspension of Milo Yiannopoulos's account gained significant media attention given his prominence in conservative circles such as Breitbart. Despite Twitter justifying its ban with Yiannopoulos's repeated violations of Twitter's rules prohibiting incitement of targeted abuse against individuals, hundreds of thousands rallied online behind the #FreeMilo hashtag.[40]

Other approaches to detecting hateful and incendiary content have not seen much success either. Word filters are crude and blunt tools, and are often both underinclusive and overinclusive in identifying prohibited content. Context is ultimately key: innocuous words may be weaponized, and often so if a platform employs word filters. For instance, the far-right in Germany refashioned the term "Nafri"—initially a common term among police to describe North Africans who sexually harass or rape women—into an insult targeted against refugees as a whole.[41] Furthermore, legislative bodies and courts from various jurisdictions have criticized the use of such preemptive censorship.[42] In particular, the speculative nature of

---

38. Olivia Solon, *Underpaid and Overburdened: The Life of a Facebook Moderator*, GUARDIAN (May 25, 2017), https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content.

39. Katrin Bennhold, *Germany Acts to Tame Facebook, Learning From Its Own History of Hate*, N.Y. TIMES (May 19, 2018), https://www.nytimes.com/2018/05/19/technology/facebook-deletion-center-germany.html.

40. Abby Ohlheiser, *Just How Offensive Did Milo Yiannopoulos Have to be to Get Banned from Twitter?*, WASH. POST (July 21, 2016), https://www.washingtonpost.com/news/the-intersect/wp/2016/07/21/what-it-takes-to-get-banned-from-twitter/?utm_term=.f72f52329243.

41. Bennhold, *supra* note 39.

42. *See, e.g.*, *New York Times Co. v. United States*, 403 U.S. 713 (1971) (holding that the government's injunctions against newspapers violate the First Amendment); Observer & Guardian v. United Kingdom, App. No. 13585/88, 14 Eur. Ct. H.R. 153 (1991) (holding that freedom of expression principles prevented the government from filing an injunction to prevent the publication of a novel); Directive 2000/31/EC, 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 (Directive on Electronic Commerce), art. 15, 2000 O.J. (L 178) 13, 13 (prohibit-

preemptively removing content and the poor track record of governmental abuse render it too bitter a pill to swallow.

Given such challenges, it is unsurprising that Facebook's internal systems only successfully flag 38% of hate speech on its platform, in stark contrast to the 96% of adult nudity and 99.5% of terrorist content flagged.[43] As such, if legislators persist in making social media companies strictly liable for all prohibited content on their platforms, then these companies would naturally prioritize removing all potentially hateful and incendiary content over attempting to adjudicate whether the content actually meets the threshold. The users and their freedom of expression will suffer as a consequence. Human rights organizations, including the Human Rights Watch, have fought back against such an approach, criticizing Germany's recent law forcing social media platforms to censor prohibited speech.[44]

One counterargument is that some large social media platforms continue to thrive despite onerous obligations to sieve out prohibited content. WeChat,[45] Sina Weibo,[46] and Baidu[47] are hugely popular in China, but are also subject to

---

ing E.U. Member States from imposing a "general obligation to monitor") [hereinafter Directive on Electronic Commerce]; Case C-360/10, Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlof NV, 2012 E.C.R. I-0000 (ruling that social networks cannot be required to monitor activities on their site to prevent copyright infringement).

43. Frenkel, *supra* note 7. R

44. Letter from Civil and Human Rights Organizations and Industry Bodies, to European Commission, Germany's Draft Network Enforcement Law Is A Threat to Freedom of Expression, Established EU Law and the Goals of the Commission's DSM Strategy—The Commission Must Take Action (May 22, 2017), https://edri.org/files/201705-letter-germany-network-enforcement-law.pdf; *Germany: Flawed Social Media Law*, HUMAN RIGHTS WATCH (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law; *Proposed German Legislation Threatens Free Expression Around the World*, GLOBAL NETWORK INITIATIVE (Apr. 20, 2017), https://globalnetworkinitiative.org/proposed-german-legislation-threatens-free-expression-around-the-world.

45. WeChat is a Chinese application with about 902 million users that allows users to message, post content, and make mobile payment. There are about 38 billion messages posted on WeChat every day. Shannon Liao, *How WeChat Came to Rule China*, VERGE (Feb. 1, 2018), https://www.theverge.com/2018/2/1/16721230/wechat-china-app-mini-programs-messaging-electronic-id-system.

46. Sina Weibo is a popular Chinese microblogging website.

47. Baidu is the dominant Chinese local search engine.

strict liability laws regarding prohibited content on their platforms. If anything, these examples highlight the reasons why strict liability should not be favored. These companies rely on extensive word filters and are unafraid of curtailing freedom of expression and online discourse, thus obviating the need for a large army of moderators. For instance, if one searches "Tiananmen" in Chinese on Baidu, no results will link to the pro-democracy protests in 1989.[48] Another example that has gained notoriety is the ban on all searches regarding Winnie the Pooh, which was used as an oblique and comical reference to Chinese President Xi Jinping.[49] With strict liability, the potential for abuse is too great to bear.

C. *Third Option: Liability for Social Media Platforms Under Certain Circumstances*

The third and final option is to hold social media companies liable under certain circumstances. Social media's role is facilitation, not publication, and such an approach recognizes the intermediary role and responsibility that these platforms have.

For instance, Brazil and Chile only impose liability on platforms for non-compliance with a court order to remove prohibited speech.[50] The 2000 European Union e-Commerce Directive requires social media platforms to remove or disable access to prohibited speech once they have actual knowledge or awareness of illegal activities.[51] The issuance of a court order for take-down constitutes actual knowledge. Argentina is

48. Simon Denyer, *China's Scary Lesson to the World: Censoring the Internet Works*, WASH. POST (May 23, 2016), https://www.washingtonpost.com/world/asia_pacific/chinas-scary-lesson-to-the-world-censoring-the-internet-works/2016/05/23/413afe78-fff3-11e5-8bb1-f124a43f84dc_story.html?utm_term=.f90149752ca9.

49. Stephen McDonell, *Why China Censors Banned Winnie the Pooh*, BBC (July 17, 2017), https://www.bbc.com/news/blogs-china-blog-40627855.

50. For Brazil, see U.N. Doc. A/HRC/17/27, *supra* note 12, ¶ 43. For Chile, see Jeremy Malcolm, *Will Big Content Derail Argentina's New Intermediary Law?*, ELECTRONIC FRONTIER FOUNDATION (Mar. 27, 2018), https://www.eff.org/deeplinks/2018/03/will-big-content-derail-argentinas-new-intermediary-law; U.N. Doc. A/HRC/17/27, *supra* note 12, ¶ 43.

51. Directive 2000/31/EC, 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 (Directive on Electronic Commerce), 2000 O.J. (L 178) 1, 6.

currently debating a new law in line with the above.[52] Apart from notice from a court order, social media platforms could rely on user notifications that flag certain controversial posts. However, they should not be held liable for not removing user-flagged content, especially considering the vast amount of content that social media platforms review. YouTube revealed that users reported nearly 10 million videos from April to June 2018 for potentially violating its community guidelines.[53] Further, this does not obviate the main difficulty of determining whether it is prohibited content. Social media platforms could choose to remove content flagged by a minimum threshold of users, but such action would dilute the freedom of expression by placing individuals at the mercy of the "heckler's veto."[54]

However, it is unclear how this option is superior to the no liability approach. Given the lengthy process of seeking a court order, the social media post may have already gone viral, with its hateful and incendiary content threatening imminent violence upon individuals or groups. Platforms may mitigate this by setting up expedited procedures for prohibited content take-down orders, but this does not sidestep concerns about how governments and social media companies can identify such potential prohibited content in a timely manner.

IV. Evaluation

Each approach has its distinct advantages and drawbacks. To determine which option governments ought to adopt, this paper proposes three evaluative criteria: accountability, sustainability, and legitimacy.

Accountability is essential in the final analysis because of the intricate role that social media platforms play in the actualization of free speech and expression. It may be a virtual space, but that does not subtract from its role as a public space for free expression and meaningful discourse. One enduring quality of free expression is that it is not bound to any medium, physical or otherwise. If social media platforms, by way of popularity and socialization, become a crucial conduit for speech

52. Malcolm, *supra* note 50.

53. Frenkel, *supra* note 7. R

54. Vajnai v. Hungary, App. No. 33629/06, HUDOC, ¶ 57 (July 8, 2008), https://hudoc.echr.coe.int/eng#{%22documentcollectionid2%22:[%22GRANDCHAMBER%22,%22CHAMBER%22]}.

and expression, then governments must take up the mantle of accountability to manage any potential violations of rights which are entangled with these social media platforms.

In the third option of limited liability, governments must request a take-down order from the courts. The courts are in turn expected to provide clear justifications for their decisions. Not only can the public and civil society scrutinize these decisions, but the decisions can also face challenge in appellate courts for violation of a constitutional provision, or in regional and international courts for violation of human rights treaties. For instance, the Indian Supreme Court struck down a vague law governing online expression that the government used to prosecute people when they legitimately exercised their right to free speech online.[55]

In contrast, social media platforms have no legal obligation to explain themselves when removing content, leaving users in the dark about when or why their posts are taken down.[56] Moreover, due process is compromised if users have no opportunity to appeal take-down requests.[57] In December 2017, Facebook deleted the accounts of Chechen Republic leader Ramzan Kadyrov on the basis that the U.S. Treasury's Office of Foreign Assets Control added Kadyrov to its sanctions list.[58] However, critics have pointed out that many other individuals on the same sanctions list remain active on Facebook.[59] Facebook has also refused to respond to any queries regarding why Kadyrov's accounts were deleted.[60] Since

---

55. Shreya v. Union of India, AIR 2015 SC 1523, para. 119 (India); *India: Historic Supreme Court Ruling Upholds Online Freedom of Expression*, AMNESTY INT'L (Mar. 24, 2015), https://www.amnesty.org/en/latest/news/2015/03/india-supreme-court-upholds-online-freedom-of-expression.

56. Mutuma Ruteere (Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance), *Rep. of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance, Mutuma Ruteere*, ¶ 53, U.N. Doc A/HRC/26/49 (May 6, 2014).

57. MACKINNON, *supra* note 36, at 40.

58. Kalev Leetaru, *Facebook's Deletion of Ramzan Kadyrov and Who Controls the Web?*, FORBES (Dec. 29, 2017, 8:28 PM), https://www.forbes.com/sites/kalevleetaru/2017/12/29/facebooks-deletion-of-ramzan-kadyrov-and-who-controls-the-web/#7db02f9f6dc8.

59. *Id.*

60. Masha Gessen, *How Chechnya's Leader Got Banned from Facebook and Instagram*, NEW YORKER (Jan. 26, 2018), https://www.newyorker.com/news/

Facebook's decision is not publicly available nor subject to appeal, Kadyrov's accounts remain banned on tenuous grounds. Although Facebook has mooted the possibility of having a Facebook "Supreme Court" to make calls on contested content moderation decisions,[61] it is unlikely that a body unaccountable to the public will be trusted to make the correct judgments on behalf of society.

Legitimacy is fundamental in generating respect for the rules and decisions surrounding prohibited content. A take-down request that lacks legitimacy only empowers the aggrieved party to claim that he or she was subject to the heckler's veto or to the whims of a single human moderator. Forcing moderators to decide whether a social media post contains prohibited content within seconds will not win votes of confidence from the public. Reports demonstrate that human moderators have to sit in front of a computer and choose within ten seconds if the post should be removed. However, determining whether speech should be prohibited is a determination even courts find challenging, as they are highly fact-sensitive.[62] More crucially, when a court decides that a particular post has crossed the line, the legitimacy that underlies the take-down request amplifies the strong message that the society stands behind the removal of certain posts because they are harmful to the community.

Sustainability is another key concern in the final analysis, as any feasible solution must account for the ever-growing use of social media and the increasing complexity of determining prohibited speech. An approach that only works today will not suffice in the long run. In this respect, forcing social media platforms to take on the responsibility of determining prohibited content will result in overbroad censorship, transforming social media platforms from public spaces of free expression into heavily controlled zones of inhibited discourse. The diffi-

---

our-columnists/how-chechnyas-leader-got-banned-from-facebook-and-instagram.

61. Ezra Klein, *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next*, VOX (Apr. 2, 2018, 6:00 AM), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge; Kate Klonick & Thomas Kadri, *How to Make Facebook's 'Supreme Court' Work*, N.Y. TIMES (Nov. 17, 2018), https://www.nytimes.com/2018/11/17/opinion/facebook-supreme-court-speech.html.

62. *Germany: Flawed Social Media Law*, *supra* note 44.

culty that these companies face in determining whether they should remove posts has been acknowledged by the European Union. In its 2000 e-Commerce Directive, the European Union required online platforms to act expeditiously to remove illegal content after they obtain knowledge of it.[63] However, the European Commission also qualifies that where social media platforms find it difficult to determine if a speech is illegal, they could obtain advice from competent authorities.[64] In fact, the European Commission expressly provides that E.U. member states shall not impose a general obligation on social media platforms to monitor all content.[65] This is most consistent with the third option, wherein social media companies are notified by the courts to remove certain content.

The approach that best fulfils the three criteria is the third: liability for social media platforms under certain circumstances. However, this paper is also cognizant of the limitations of that option. If social media platforms are only required to take down content when a court order is issued, prohibited speech would remain rife online. By the time a court order is issued, violence might have erupted on the streets. Ultimately, reliance on a purely legal framework is insufficient. The legal framework must be supplemented by a partnership model between governments and social media platforms to moderate speech online. Some governments have already pursued this option by setting up co-monitoring units. After the Paris terror attack in 2015, France established a six-month partnership with Facebook to focus on how they can collaborate to remove prohibited content on Facebook.[66] Germany has also partnered with Facebook to counter posts containing hateful

---

63. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms*, ¶ 13, COM (2017) 555 final (Sept. 28, 2017) (citing Directive on Electronic Commerce, *supra* note 51, at 6).

64. European Commission, *supra* note 63, ¶ 13.

65. Directive on Electronic Commerce, *supra* note 42, at 13.

66. Makena Kelly, *Facebook Will Allow French Regulators to Monitor Content Moderation Processes*, VERGE (Nov. 12, 2018), https://www.theverge.com/2018/11/12/18089012/facebook-france-emmanuel-macron-hate-speech; Mark Scott & Zachary Young, *France and Facebook Announce Partnership Against Online Hate Speech*, POLITICO (Nov. 12, 2018), https://www.politico.eu/article/emmanuel-macron-mark-zuckberg-paris-hate-speech-igf.

statements against refugees.[67] This partnership model should be replicated on a permanent basis, with social media platforms taking the initiative to cooperate with governments on sieving out prohibited speech. Otherwise, when faced with the increasing risk of violence that hateful social media content instigates, governments are likely to follow in Germany's footsteps and simply force these platforms to take on more liability in regulating content.

V. CONCLUSION

The question of when governments should hold social media companies liable for prohibited speech on their platforms is a thorny one that requires careful consideration of the practicalities involved. Harsh penalties may inadvertently curtail individuals' freedom of expression, but a lenient punishment would allow prohibited content to proliferate, fomenting distrust and sparking violence between different socioeconomic groups. This paper proposes that the best balance is achieved when governments and social media platforms jointly tackle the challenges of identifying and removing prohibited content in an accountable, legitimate, and sustainable way.

67. Amar Toor, *Facebook Will Work with Germany to Combat Anti-Refugee Hate Speech*, Verge (Sept. 15, 2015), https://www.theverge.com/2015/9/15/9329119/facebook-germany-hate-speech-xenophobia-migrant-refugee.