Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
NOV. 2022 VOL.1 NO.4

Privacy, Free Speech and Content Moderation: A Literature Review and Constitutional Framework Analysis

Yi Wu[1]

[1] UC Berkeley, CA, US

Correspondence: Yi Wu, UC Berkeley, CA, US.

doi:10.56397/IST.2022.11.04

Abstract

Content moderation is one of the lifeblood of Internet platform companies. Concealment and innovation coexist. The United States and the European Union are leading the world in the field of human rights protection in content moderation. Both the text of the Constitution and the jurisprudence have formed a relatively complete framework of protection. Public and private sectors, online and offline communities, and the balance between ex-ante and ex-post, are at the intersection of content moderation with privacy and free speech. We need to further research exploring the design of the online moderation model, which will balance the arguments around policy, human rights law, and the need to make online spaces safer for a world-wide diverse population.

Keywords: privacy, free speech, content moderation, constitutional framework

1. Introduction

The rise of online media is one of the most influential inventions on the Internet in the 21st century. In the US, 72% of the public uses some type of social media in 2021, compared to 5% in 2005.[1]

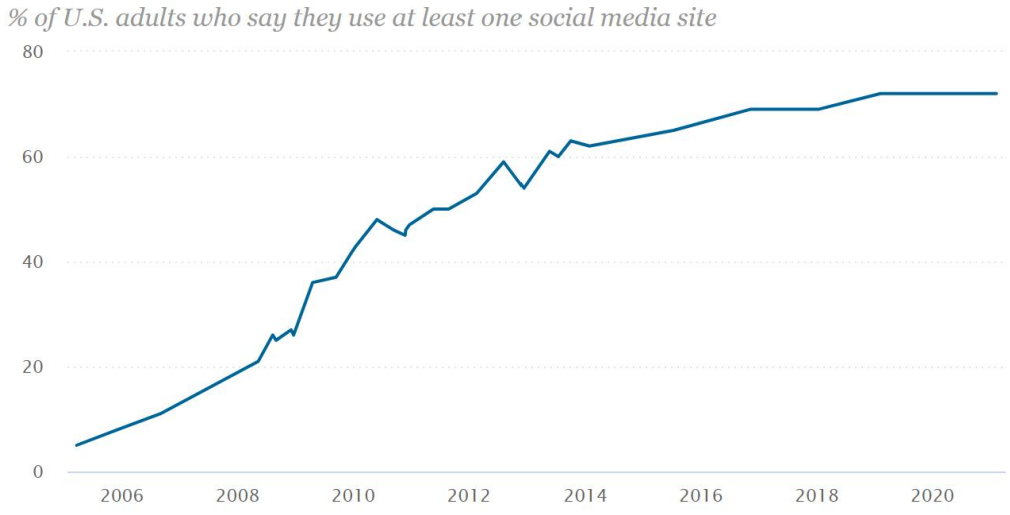

Figure 1. Public uses of social media in U.S.

For many users, social media is part of their daily routine. 70% Facebook users and around 60% Instagram and Snapchat users visit these sites at least once a day.

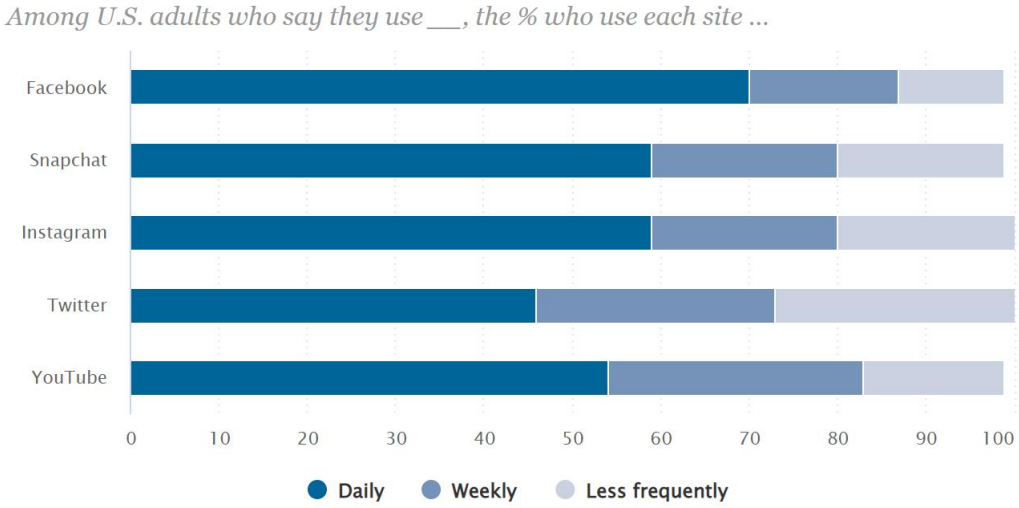

Figure 2. Social media daily uses in U.S.

Majorities of 18- to 29-year-olds say they use Instagram or Snapchat and about half say they use TikTok, with those on the younger end of this cohort – ages 18 to 24 – being especially likely to report using Instagram (76%), Snapchat (75%) or TikTok (55%). These shares stand in stark contrast to those in older age groups. For instance, while 65% of adults ages 18 to 29 say they use Snapchat, just 2% of those 65 and older report using the app – a difference of 63 percentage points.

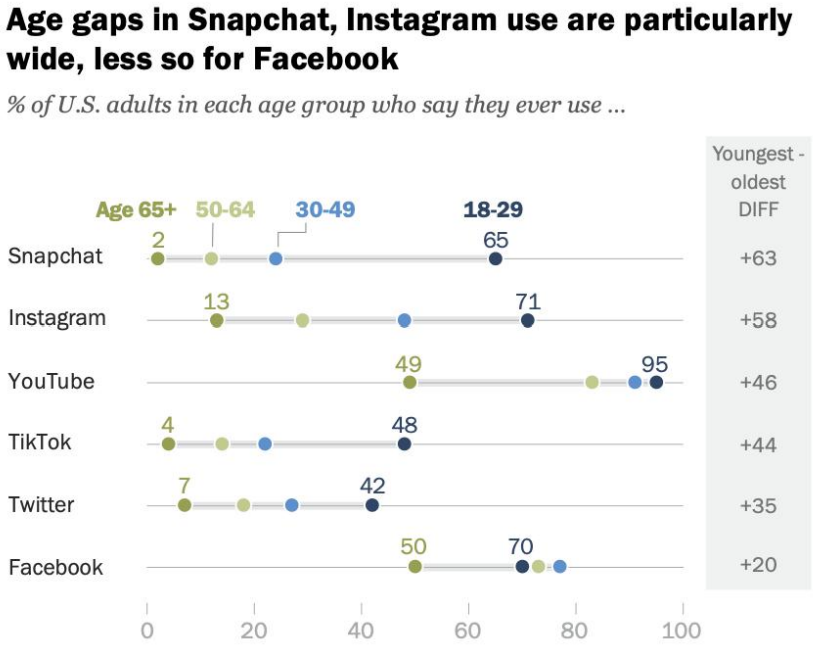

Figure 3. Age distribution of social media uses in U.S.

In order to meet the state regulatory requirements and the healthy development of the platform itself, content moderation is an indispensable step. According to TMR Report, the Content Moderation Solutions Market to Cross US$ 32 Bn by 2031.[2] Moderation can be defined as "the screening, evaluation, categorization, approval or removal/hiding of online content according to relevant communications and publishing policies. It seeks to support and enforce positive communications behavior online, and to minimize aggression and anti-social

behavior". (Terry Flew & et al., 2019)

Accordingly, laws and regulations in the field of content censorship are gradually improving, with the United States and the European Union taking the leading position in the world.

At the last century, the US and EU built the system of online intermediaries' liability based on a liberal regulatory approach. When the US Congress passed Section 230 of the Communication Decency Act ('CDA') in 1996[3], the primary aim was to encourage the sharing of free expression and development of the digital environment. (Kate Klonick, 2018) In order to achieve this objective, the choice was to exempt computer services from liability for merely conveying third-party content.

Before the adoption of the CDA, some cases had already made clear how online intermediaries would have been subject to a broad and unpredictable range of cases concerning their liability for editing third-party content. [4]

Since this risk would have slowed down the development of new digital services in the aftermath of the Internet, online intermediaries have been encouraged to grow and develop their business under the protection of the Good Samaritan rule.[5]

The Digital Millennium Copyright Act ('DMCA') introduced in 1997 allows online intermediaries not to be held liable for hosting unauthorized copyright works.[6] Nevertheless, unlike the CDA, the DMCA does not provide an absolute exemption but shield online intermediaries from liability according to certain conditions.[7]

In the EU, the e-Commerce Directive, adopted in 2000, exempts hosting providers (e.g., social network or search engine) from liability for third-party content, provided that they remove or disable online content once they become aware of its unlawful nature.[8]

The General Data Protection Regulation (GDPR), approved in 2016, is the toughest privacy and security law in the world. Though it was drafted and passed by the European Union (EU), it imposes obligations onto organizations anywhere, so long as they target or collect data related to people in the EU.[9]

There are some differences between the European Union and the United States, such as the right to free speech. While, in the US, the Internet and social media still benefit from the frame coming from the traditional liberal metaphor of the free marketplace of ideas as a safeguard for democracy, in Europe, the protection of freedom of expression online does not enjoy the same degree of protection. [10]In the European framework, the right to freedom of expression is subject to a multilevel balancing with other rights enshrined in the Charter of Fundamental Rights of the European Union ('Charter')[11], the European Convention of Human Rights ('Convention')[12], and national constitutions.

The other important framework in the field of content moderation is GIFCT. Global Internet Forum to Counter Terrorism, is an NGO designed to prevent terrorists and violent extremists from exploiting digital platforms. Founded by Facebook, Microsoft, Twitter, and YouTube in 2017. Since 2017, GIFCT's membership has expanded beyond the founding companies to include over a dozen diverse platforms committed to cross-industry efforts to counter the spread of terrorist and violent extremist content online.[13]

Till now, the U.S. and Europe are in the throes of a fundamental intermediary liability legislative fight: who deserves safeguarding, what are the major threats, and can government rewrite the rules without pulling the plug on the internet as we know it? This literature view examines the limits of content moderation on human right to privacy and free speech under the constitutional framework.

2. Privacy and Content Moderation

Privacy and security, it's difficult to provide one without violating the other. For example, scanning a user's inbox for potentially malicious messages seems to imply access to all safe ones as well. (Seth Frey. Maarten W.Bos & Robert W. Sumner, 2017) However, privacy is different from security. For example, although companies may promise to protect users' information from access by third parties, malicious players (e.g., hackers; disgruntled employees) may obtain that information anyway, which does not imply a breach of privacy, but rather a breach of security. (Elissa M. Redmiles, Jessica Bodford & Lindsay Blackwell, 2019)

The study of Pariser gives a warning: in the unmapped territory of online privacy, the onus is on consumers to maintain an enforceable barrier between public and private information. (Pariser, E., 2011)

Different platforms have different scope and form of privacy protection. Reddit does not encourage participation with one's real name as a privacy-protecting measure (Nicholas Proferes, Naiyan Jone, Sarah Gilbert, Casey Fiesler4 & Michael Zimmer, 2021). In a case study of Online Patient Communities, privacy violation is being detected. Although more information regarding a member's condition and situation helps others relate to the member's issues and provide appropriate support. However, sometimes this is taken too far. Violations to privacy may occur when a member shares excessive personal health information: members often ask questions such as "how was this diagnosed," "how long you have had the problem," and "what kinds of treatment have you had."

(Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna1, Sami Safadi & Fouad Chebib, 2021)

Tim analyzed over 80,000 health-related web pages and determined that 90 percent leaked user information to outside parties. (Tim Libert, 2014) 70 percent of health-related websites have addresses which contain information on specific symptoms, treatments, and diseases.

A Facebook user said, in once investigation, "I was shocked to keep seeing my privacy settings changed from who can see my posts. 'Friends' I set, but it says 'Public!' This reset or someone is resetting it! I need to find out why. I usually ask my daughters first. Then I will be contacting Facebook." These findings echo prior work that expectation setting is key for avoiding loss of trust and feelings of privacy violation (Rao, A., Schaub, F., Sadeh, N.; Acquisti, A.& Kang, R., 2016).

While a single Facebook programmer may only care about optimizing the site for the user's preferences, the market ensures that (privacy policy or not) users can be observed, profiled, and most importantly, recognized on subsequent visits. (Samuels & Mark Gregory, 2012)

A lot of personal information is mishandled behind our browser's digital veil. The matrix of engineers, market demographers, and data aggregators are contracted third parties to our trusted hosts; they are not contractually bound to the same privacy standards that Facebook acquiesced to in 2010. The right to privacy aegis, which many Americans imagine protects them, is actually a loose net of judicial rulings and corporate best practices.

Although freedom of expression and confidentiality of communications are primary considerations and users of telecommunications and internet services must have a guarantee that their own privacy and freedom of expression will be respected, such guarantee cannot be absolute and must yield on occasion to other legitimate imperatives, such as the prevention of disorder or crime or the protection of the rights and freedoms of others. A case in KU v. Finland, the Court observed:[14]

Although freedom of expression and confidentiality of communications are primary considerations and users of telecommunications and internet services must have a guarantee that their own privacy and freedom of expression will be respected, such guarantee cannot be absolute and must yield on occasion to other legitimate imperatives, such as the prevention of disorder or crime or the protection of the rights and freedoms of others. Without prejudice to the question whether the conduct of the person who placed the offending advertisement on the internet can attract the protection of Articles 8 and 10, having regard to its reprehensible nature, it is nonetheless the task of the legislator to provide the framework for reconciling the various claims which compete for protection in this context. Such framework was not, however, in place at the material time, with the result that Finland's positive obligation with respect to the applicant could not be discharged.

The fact that content moderation cannot be completely automated is not limited by technology, but by the nature of the content. As a way of supervised learning, machine learning cannot replace human beings in its prediction and detection ability.

While the activities of pre-moderation like prioritization, delisting and geo-blocking are usually automated, post-moderation is usually the result of a mix between automated and human resources. (Sarah T. Roberts, 2017)

3. Freedom of Speech and Content Moderation

The United States stands out as a bastion of freedom of expression within the new digital ecosystem. (L. Bollinger, 1986) It is the result of the broader constitutional protection afforded by the First Amendment to the US Constitution. It enjoys a broader scope of protection than that in Europe, where free speech is recognized as a fundamental right whose protection needs to be balanced with other constitutional interests.

The two main theories of interpreting the First Amendment in the US are revolving around the libertarian notion of a marketplace of ideas and the republican one of self-government. (B. Petkova, 2019) This stance is apparent from the value frame adopted by the Supreme Court, which has since the very first judgments in this area been inclined to exalt the unprecedented libertarian dimension to the internet. From the US Supreme Court perspective, the internet has offered new fora and spaces for the exercise of freedom of expression from the outset, which must be viewed through different lenses and with reference to different categories from those applied to traditional media. This constituted the basis for the choice, in Reno v. ACLU30,[15] to borrow the metaphor of the 'free marketplace of ideas' from the renowned dissenting opinion of Justice Holmes[16], the enduring relevance of which could be called into question in the light of the changes to the internet over the early years of its existence.

In practice, lager internet platforms have formed their own cognition and protection of free speech in practice. Reddit provides an interesting site of study into content moderation issues due to a culture of debate over whether free speech is a principal tenet of the platform. (Adi Robertson, 2015) Both left-leaning and right-leaning users, for example, used statements decrying both hate speech and censorship and highlighted concerns with how the Reddit quarantine policy was implemented. Instead, some scholars argue that these

strategies are employed as a defense of a user's legitimate participation on Reddit. While previous work has examined the use of free speech discourse as a defense against ego or expressive threat (II White, H Mark & Christian S Crandall, 2017), further exploration is needed into why the specific strategies of censorship vs. consistency are applied in the context of online discussion.

Policy experts have posed bilateral arguments around the notions of freedom of speech and a platform's responsibility in restricting and moderating hateful communication.

Aswad's analysis outlines several challenges associated with deriving meaningful online content moderation policies while also aligning them with international human rights law. (Evelyn Mary Aswad, 2018) McDonald et al.'s study also laid out the challenges in online governance of extremist content, including lack of clear directive and the inability of moderation algorithms to distinguish different types of extremism. (Stuart Macdonald, Sara Giro Correia & Amy-Louise Watkin, 2019)

Particularly in the light of recent censorship of white nationalism on Facebook[17], hate groups might be quickly adapting and moving their online operations in alternative platforms championing free speech. (Shruti Phadke, Tanushree Mitra, 2020)

While political correctness generally refers to discursive strategies or principles of avoiding utterances and actions that could offend or marginalize particular groups of people (largely corresponding with protected characteristics against hate speech in Facebook's Community Standards), it has recently become a "spurious construct" (Fairclough & Norman, 2003) or a metapragmatic label for an ideological other often associated with imposing censorship and limiting freedom of speech, especially in the right-wing conservative circles. (Ondřej Procházka, 2019) It's difficult it is to accurately define "hateful speech" or the limits of "freedom of expression." (Minna Ruckenstein, Linda Lisa & Maria Turunen, 2020)

Different free speech is examined by liberals and conservatives offline in the debate over campus free speech. (Jonathan Friedman, 2019) Platforms have worked hard to maintain a public image of neutrality (Roberts ST, 2018), with moderators talking about freedom of expression to explain their position, thereby signaling the avoidance of excessive regulation. Yet, they also understand, through hands-on experience, the tensions and difficulties of limiting online content. It might not be easy to distinguish whether the message should be evaluated as a statement, an intention or a threat. Depending on how its motive is interpreted, the message is categorized as either illicit or appropriate—as it is legitimate to share opinions in a neutral way.

From the user's point of view, it is difficult to detect their privacy is violated, more cannot really exercise rights. In a survey conducted to research Facebook (Kristen Vaccaro, Christian Sanding & Karrie Karahalios, 2020), While many participants mention their agreement with the basic goals (if you do something wrong then you should have privileges taken away from you), many others fundamentally argue that Facebook should not moderate content. Many of these participants refer to basic principles, like freedom of speech to argue that their content should not be removed:

Facebook has a right to remove offensive information, but they don't have to right to suppress freedom of speech, including if it is different from their views or especially their political leanings.

And while many appreciate that Facebook was attempting to curb the spread of misinformation, others argue that Facebook should have other priorities than the spread of misinformation.

I think it's stupid. You need to focus more on the cyberbullying and catfishing or that misinformation does not really harm other users: If it isn't an illegal action activity taking place on the page, I think there's no harm to anyone nor Facebook.

4. Content Moderation in the Constitutional Framework

Constitutional Law usually refers to rights granted by the U.S. Constitution. Cases often involve the Bill of Rights, or respective rights of federal and state governments. Its provisions are the limit to the coercive power of the State or as a source of positive obligation for public actors to protect constitutional rights and liberties.

The scope of constitutional rights allows private actors to claim the respect of their rights only vis-à-vis public actors. In the algorithmic society, instead, an equally important and pernicious threat for individuals come from those private actors which develop algorithms according to their ethical, economic and self-regulatory frameworks. While, in the past, the threats for individual rights were linked with State actions, today, democratic States deal with the issue of limiting the exercise of freedoms (or powers) exercised by private actors in the digital environment.

In the civil sphere, one's right of free speech must be balanced against their duty not to cause harm to others. In legal settings, this balance is defined by laws (Waldron, J., 2012), or courts' interpretation of those laws. In the social media world, this balance is defined through the standards and policies that social media companies create. In the new digital ecosystem, the Supreme Court embraced the metaphor of the free marketplace of ideas, and in

fact exalted it by identifying the internet as a special domain in which this free exchange could be considerably expanded.

A growing body of literature recognizes that non-state actors, like Instagram, have become 'the new governors' (Kate Klonick & Tarleton Gillespie, 2018) of the digital age. (Lawrence Lessig, 1999; James Grimmelmann, 2005; Colin Scott, 2004) Contractual terms of service arguably function as types of constitutional documents in the way that they establish the power of platform owners to regulate user-generated content and set standards of 'appropriate' behavior. (Nicolas Suzor, 2018; Jessica Anderson & et al, 2016) As Facebook CEO Mark Zuckerberg acknowledged in 2009, 'our terms aren't just a document that protects our rights; it's the governing document for how the service is used by everyone across the world'. (Adweek Staff, 2009) (Kyle Langvardt, 2018) Yet terms of service make poor constitutional documents.

Unlike traditional constitutions, this contractual bargain affords platform owners 'complete discretion to control how the network works and how it used' by users, who are the subjects of platform governance. (Kate Crawford & Tarleton Gillespie, 2016) Instagram's Terms of Use, for instance, states that 'we can remove any content or information you share on the Service if we believe that it violates these Terms of Use, our policies (including our Instagram Community Guidelines), or we are permitted or required to do so by law'.[18]Users have little say in determining the content of terms of service, and there is little effective choice in the market – over two billion of the world's active, monthly social media users can either 'take it or leave it'. (Marjorie Heins, 2014)

Private companies may lack legitimacy for engaging in rulemaking and enforcement. For example, in the US many users believe that they are entitled to free expression in their social media behavior and are upset when their content is removed, even though the right to free speech enshrined in the First Amendment prohibits intrusions by the government, not private entities. (Tom Tyler, Matt Katsaros, Tracey Meares & Sudhir Venkatesh, 2019)

This impacts sites in several ways. First, unless users buy into rule enforcement decisions, they can seek to subvert them. One common approach is to open multiple user accounts.

Another is to operate in smaller or more private spaces to avoid notice by the companies. Further, companies may find that their approach of removing problematic content has the effect of alienating users and increasing future rule-breaking behavior. Hence, the advantages of social media companies in the arena of rulemaking and enforcement are not unlimited.

Studies of law enforcement and courts suggest that evaluations of the procedural justice of the actions of the authority will influence later behavior (Tyler, T.R., Goff, P.A. & MacCoun, R.J., 2015). Its specific performance is pre-examination and post-examination. Users are critical pieces of the content moderation puzzle since social media also rely on users to flag or, generally, report content. (Kate Crawford & Tarleton Gillespie, 2016) The platform needs content notice to explain to users how their content is processed and according to which conditions. Moderation of content is also autonomously performed ex-ante by automated means, for instance, to tackle extreme content like terrorist videos when uploaded. According to the current system, the platforms mitigate the fragmentation to establish forms with mandatory information. For instance, the notice could not include the URL to identify the content at stake or do not explain the issue at stake.[19]

While this extension might appear to be straightforward, there are two potential uncertainties.

First, private companies may be viewed as inappropriate entities for decision makings about what users are entitled to do. They may be viewed as lacking the training and expertise associated with authorities like judges, and, aside from their control over user accounts, the nature of their authority to make such decisions is sometimes disputed by users. Meanwhile, they are not legal authorities from the user's perspective.

Second, procedural justice is in doubt. Its effectiveness mainly depends upon users identifying with the group, organization or community the authority represents (Schultz, M.F., 2006). Users tend to not feel any type of identification with the proprietors of such sites. After all, referees and athletes should come from different entities because they represent different interests. Mixed together, it is difficult to let a person do not have easy to misunderstand the association. Social media sites are communities of a type, and they may be what communities look like in the modern world (Gruzd, A., Jacobson, J., Wellman, B. & Mai, P., 2016). The question is whether they support the type of value-based self-regulation found with legal authorities.

5. Discussion

Opacity is one of the causes of possible violations of constitutional rights. Despite the fundamental role of social media platforms in establishing the standard of free speech and shaping democratic culture on a global scale (Marvin Ammori, 2014), the information provided by these companies about content moderation is opaque or law-less threatening the rule of law. (Nicolas Suzor, 2019)

Facebook has been in the firing line for years. Facebook has also faced significant criticism (and popular press

coverage) for its policy decisions, from the early outcry over its research policy (Alex Hern, 2014) to more recent fact checking (Dave Lee., 2019), ad placement (Bill Chappell, 2020), and privacy policies (Emily Stewart, 2018). In 2018, Facebook has published a "Draft Charter" of its 40-person oversight board explaining the commitments of the board such as transparency and motivation of the decisions.[20]

From a policy perspective, lack of appropriate transparency measures regarding data collection and AI decision making enforced upon companies may be contributing to disparate impact. Ultimately, algorithmic bias cannot be mitigated without algorithmic transparency. (Akshat Pande & Aylin Caliskan, 2021) In the case of complex deep learning algorithms, creating transparency is much more difficult. It is not the reluctance of a company or organization or the complexity of correlations that prevents openness, but the nature and design of the algorithms themselves. (Bert Heinrichs, 2022) At least the first point is hard to prove. What's more challenging is service providers can claim commercial secret protection against outsiders from abusing or studying their technology. (Darren Stevenson, 2014)

There is a common issue in both public and private sectors: Can governments enforce rules in ways that promote citizen/user acceptance, and enhance their willingness to self-regulate in the future as well?

We need to further research exploring the design of online censorship and moderation model, which carefully balances the arguments around policy, human rights law, and the need to make online spaces safer for a diverse population.

Content moderation decisions make fundamental choices about what speech should be supported or silenced. Many have argued that social media has become an important "public square," (Jeffrey S Juris, 2012) despite private ownership (Bill Sherman, 2011). However, many users have concerns about free speech on these platforms. And if social media platforms constitute a public square, then the public has an interest in ensuring that important speech is not being silenced.

Multidimensional platforms such as social media sites present more safety challenges for users as well as for developers and designers, who must balance the needs of potentially competing dimensions. For example, how do you balance privacy with the desire to share personal information for the purpose of community and relationship-building? Some scholars have creatively proposed the concept of perceived fairness. (Kaibel, Chris, Irmela Koch-Bayram, Torsten Biemann & Max Muhlenbock, 2019) This puts higher demands on large platforms.

The perceptions of online privacy are not exclusively related to digital privacy; rather, they are intertwined with users' online and offline communities. Whether content moderation is invading inevitably or protecting appropriately privacy and free speech naturally extends to a much broader scope. This is also the significance of discussing content and human rights protection in the framework of the Constitution.

References

Terry Flew & et al., (2019). Internet Regulation as Media Policy: Re-thinking the Question of Digital Communication Platform Governance. *Journal of Digital Media & Policy, 10*(1), 33, 40.

Kate Klonick, (2018). The New Governors: The People, Rules, and Processes Governing On-line Speech *Harvard Law Review, 131,* 1598.

Seth Frey. Maarten W.Bos, Robert W. Sumner, (2017). Can you moderate an unreadable message? Blind content moderation via human computation. *Human Computation, 4*(1), 78-106.

Elissa M. Redmiles, Jessica Bodford, Lindsay Blackwell, (2019). I Just Want to Feel Safe: A Diary Study of Safety Perceptions on Social Media, *ICWSM*.

Pariser, E., (2011). *The filter bubble: What the Internet is hiding from you*. New York: Penguin Press.

Nicholas Proferes, Naiyan Jone, Sarah Gilbert, Casey Fiesler4, and Michael Zimmer, (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society.*

Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna1, Sami Safadi, Fouad Chebib, (2021). Successful Moderation in Online Patient Communities: Inductive Case Study. *Journal of Internet Research*.

Tim Libert, (2014). Health Privacy Online: Patients at Risk. Open Technology Institute.

Rao, A., Schaub, F., Sadeh, N.; Acquisti, A., and Kang, R., (2016). Expecting the unexpected: Understanding mismatched privacy expectations online, *Symposium on Usable Privacy and Security (SOUPS), 4*(2).

Samuels, Mark Gregory, (2012). Review: The Filter Bubble: What the Internet is Hiding from You by Eli Pariser. Interactions: UCLA *Journal of Education and Information Studies, 8*(2).

Sarah T. Roberts, (2017). Content Moderation, in Laurie A. Schintler and Connie L. McNeely (Eds), *Encyclopedia of Big Data (Springer)*.

L. Bollinger, (1986). *The Tolerant Society*. Freedom of Speech and Extremist Speech in America, (OUP).

B. Petkova, (2019). Privacy as Europe's First Amendment, in this special issue.

Adi Robertson, (2015). Was Reddit always about free speech? Yes and no. The Verge. https://www.theverge.com/2015/7/15/8964995/reddit-%20free-speech-history.

II White, H Mark, and Christian S Crandall, (2017). Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology.*

Evelyn Mary Aswad, (2018). The Future of Freedom of Expression Online. *Duke L. & Tech. Rev., 17*, 26.

Stuart Macdonald, Sara Giro Correia, and Amy-Louise Watkin, (2019). Regulating terrorist content on social media: automation and the rule of law. *International Journal of Law in Context, 15*(2), 183-197.

Shruti Phadke, Tanushree Mitra, (2020). Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups. CHI 2020, paper. 7.

Fairclough, Norman, (2003). Political Correctness: The Politics of Culture and Language. *Discourse & Society ,14*(1), 25.

Ondřej Procházka, (2019). Making Sense of Facebook's Content Moderation: A Posthumanist Perspective on Communicative Competence and Internet Memes, 380-381.

Minna Ruckenstein, Linda Lisa, Maria Turunen, (2020). Re-humanizing the platform: Content moderators and the logic of care. SAGE. *New Media & Society, 22*(6), 1026-1042.

Jonathan Friedman, (2019). Chasm in the Classroom: Campus Free Speech in a Divided America. Technical report, PEN America.

Roberts ST, (2018). Digital detritus: Error and the logic of opacity in social media content moderation. *First Monday, 23*(3). Available at: https://firstmonday.org/ojs/index.php/fm/article/ view/8283/6649.

Kristen Vaccaro, Christian Sanding, Karrie Karahalios, (2020, October). At the End of the Day Facebook Does What It Wants: How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum-Comput. Interact*, 4(CS CW2), Article 167.

Waldron, J., (2012). *The harm in hate speech.* Harvard.

Kate Klonick, (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review, 131,* 1598-1602; Tarleton Gillespie, (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale University Press).

Lawrence Lessig, (1999). *Code and Other Laws of Cyberspace* (Basic Books) 220; James Grimmelmann, (2005). Regulation by Software, *Yale Law Journal*, *114*, 1719; Colin Scott, (2004). Regulation in the Age of Governance: The Rise of the Post-regulatory State in Jacint Jordana and David Levi-Faur (eds), *The Politics of Regulation: Institutions and Regulatory Reforms for the Age of Governance* (Edward Elgar Publishing), 145.

Nicolas Suzor, (2018). Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms, *Social Media + Society, 4*(3); Jessica Anderson et al, (2016, November 16). Censorship in Context: Insights from Crowdsourced Data on Social Media Censorship (Research Report, Onlinecensorship.org). <https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-second-report-censorship-in-context-pdf>; Ranking Digital Rights, (2018, April). 2018 Corporate Accountability Index (Research Report). <https://rankingdigitalrights.org/index2018/assets/static/download/RDRindex2018report.pdf>.

Adweek Staff, (2009, February 18). Facebook Reverts Terms of Service after Complaints, Adweek (online), <http://www.adweek.com/digital/facebook-reverts-terms-of-service-after-complaints/>.

Kyle Langvardt, (2018). Regulating Online Content Moderation. *Georgetown Law Journal, 106*, 1353-1357.

Kate Crawford and Tarleton Gillespie, (2016). What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint, *New Media & Society, 18*, 410, 412.

Marjorie Heins, (2014). The Brave New World of Social Media Censorship. *Harvard Law Review,* 325-326. <http://firstmonday.org/ojs/index.php/fm/article/view/8283/6649>; Facebook, Stats (31 December 2018) <https://newsroom.fb.com/company-info/>.

Tom Tyler, Matt Katsaros, Tracey Meares, Sudhir Venkatesh, (2019). Social media governance: Can companies motivate voluntary rule following behavior among their users.2019.1.

Tyler, T.R., Goff, P.A. & MacCoun, R.J., (2015). The impact of psychological science on policing in the United States. *Psychological Science in the Public Interest, 16*(3), 75-109.

Kate Crawford and Tarleton Gillespie, (2016). What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint, *New Media & Society, 18*, 410-411.

Schultz, M.F., (2006). Fear and norms and rock and roll: What jambands can tell us about persuading people to obey copyright law. *Berkeley Technology Law Journal, 21*, 651-728.

Gruzd, A., Jacobson, J., Wellman, B. & Mai, P., (2016). Understanding communities in an age of social media. *Information, Communication & Society, 19*, 1187-1193.

Marvin Ammori, (2014). The New New York Times: Free Speech Lawyering in the Age of Google and Twitter', *Harvard Law Review*, *127*, 2259.

Nicolas Suzor, (2019). *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press).

Alex Hern, (2014). Facebook T&Cs introduced research policy months after emotion study. https://www.theguardian. com/technology/2014/jul/01/facebook-data-policy-research-emotion-study.

Dave Lee., (2019). Facebook's Zuckerberg grilled over ad fact-checking policy. https://www.bbc.com/news/technology- 50152062.

Bill Chappell, (2020). FEC Commissioner Rips Facebook Over Political Ad Policy: This Will Not Do. https://www.npr.org/2020/01/09/794911246/fec-commissioner-rips-facebook-over-political-ad-policy-this-will-not-do.

Emily Stewart, (2018). Mark Zuckerberg testimony: the Facebook data privacy question he won't answer. https://www.vox.com/policy-and-politics/2018/4/11/17225518/mark-zuckerberg-testimony-facebook-privacy-settings-sharing.

Akshat Pande, Aylin Caliskan, (2021). Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithm. AIES'21. Virtual Event, USA.

Bert Heinrichs, (2022). Discrimination in the age of artificial intelligence. *AI & Society, 37*, 143-154.

Darren Stevenson, (2014). Locating Discrimination in Data-Based Systems. Open technology institute.

Jeffrey S Juris, (2012). Reflections on# Occupy Everywhere: Social Media, Public Space, and Emerging Logics of Aggregation. *American Ethnologist, 39*(2), 259-279.

Bill Sherman, (2011). Your Mayor, Your Friend: Public Officials, Social Networking, and Unmapped New Public Square. *Pace Law Review, 31*, 95.

Kaibel, Chris, Irmela Koch-Bayram, Torsten Biemann, and Max Muhlenbock, (2019). Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderation. In Academy of Management Proceedings: Academy of Management Briarcliff Manor NY 10510.

---

[1] Social Media Fact Sheet. Pew Research Center, 2021.

[2] https://www.prnewswire.com/news-releases/content-moderation-solutions-market-to-cross-us-32-bn-by-2031-tmr-report-301514155.html.

[3] Communication Decency Act, (1996). Section 230.

[4] Cubby, Inc. v CompuServe Inc. (1991). 776 F. Supp. 135 (S.D.N.Y.); Stratton Oakmont, Inc. v Prodigy Services Co. (1995). WL 323710 (N.Y. Sup. Ct. May 24).

[5] Zeran v Am. Online, Inc. (1997). 129 F.3d 327, 330 (4th Cir.). Davis S. Ardia, Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity under Section 230 of the Communications Decency Act, (2010). 43 Loyola of Los Angeles Law Review 373.

[6] Digital Millenium Copyright Act, (1997).

[7] Ibid, 17 U.S. Code Section 512(c).

[8] 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] OJ L 178/1. See Art 14.

[9] Jake Frankenfield. General Data Protection Regulation (GDPR) Definition and Meaning. https://www.investopedia.com/terms/g/general-data-protection-regulation-gdpr.asp.2020.

[10] Oreste Pollicino and Marco Bassini, Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis, in Andrej Savin and Jan Trzaskowski (Eds), *Research Handbook on EU Internet Law* (Edward Elgar 2014).

[11] Charter of Fundamental Rights of the European Union, (2012). OJ C326/12. Art 52.

[12] European Convention on Human Rights, (1950). Art 10(2).

[13] https://gifct.org/.

[14] KU v. Finland, App. no. 2872/02 (2008).

[15] Reno v. ACLU, (1997). 521 U.S. 844.

[16] Dissenting opinion Justice Holmes in the US Supreme Court case, Abrams v. United States [1919] 250 U.S. 616.

[17] Facebook. 2019. Standing Against Hate | Facebook Newsroom. https://newsroom.fb.com/news/2019/03/standing-against-hate/. (March 2019). (Accessed on 09/13/2019).

[18] Instagram Help Centre, Terms of Use, above n 23, [Content Removal and Disabling or Terminating Your Account].

[19] Case C-324/09 L'Oréal SA and Others v eBay International AG and Others, (2011). ECR I-6011, para 122.

[20] Santa Clara Principles on Transparency and Accountability in Content Moderation (2018).