REIMAGINING SOCIAL MEDIA GOVERNANCE: HARM, ACCOUNTABILITY, AND REPAIR

Sarita Schoenebeck & Lindsay Blackwell[*]

INTRODUCTION

Social media companies have attracted widespread criticism for the proliferation of harmful behaviors on their platforms. Individual users levy hate speech and harassment at their peers; state actors manipulate networks of fraudulent accounts to propagate misinformation; extremist groups leverage recommendation systems to recruit new members. While these and similar harmful behaviors are extensions of existing social phenomena and not inventions of the internet age, they are exacerbated and intensified by the specific technological affordances of social media sites, including visible network relationships, quantified social endorsement (e.g., "likes" and follows), and algorithmic feeds designed to maximize social engagement.

Because of the scale at which contemporary social media platforms operate—Facebook recently reported 1.84 billion daily active users[1]—traditional forms of social media governance, such as the appointment of volunteer moderators, have struggled to keep apace. Social media companies have attempted to address these concerns by developing formal content moderation policies and enforcement procedures, but they are not made transparent to users,

---

[*] Sarita Schoenbeck, Professor, School of Information, University of Michigan; Lindsay Blackwell, PhD Candidate, School of Information, University of Michigan.

[1] *Fourth Quarter 2020 Results Conference Call*, FACEBOOK (Jan. 27, 2021), http://s21.q4cdn.com/399680738/files/doc_financials/2020/q4/FB-Q4-2020-Conference-Call-Transcript.pdf.

both in process and outcome.[2] Scaled content moderation also requires significant human labor—typically outsourced to third-party contractors who earn relatively low wages for work that is both physically and emotionally taxing[3]—to review individual pieces of content for potential policy violations, which results in delayed response times and backlogs of lower-priority violations.

Though regulators, researchers, and practitioners alike agree that change is needed, experts disagree on best paths forward. We propose a reframing of social media governance focused on repairing harm. Repairing harm requires recognizing that harm has occurred; centering the needs of individuals and communities who experience harm; and accepting accountability for the harm, both for the specific instance of harm and its root causes.

We first review prominent paradigms for the regulation of online behavior, from the 1980s through the early 2020s. Then, we discuss common categories of harm experienced on or created by social media platforms, including the consequences of inadequate platform governance. Drawing on principles of retributive, restorative, and transformative justice, we propose social media governance frameworks for better addressing those harms. We argue that, although punishment is sometimes necessary, a solely punitive model of governance is insufficient for encouraging compliance or for deterring future harm. We conclude with several

---

[2] *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, https://santaclaraprinciples.org (last visited Jan. 10, 2021) [hereinafter *The Santa Clara Principles*]; JILLIAN C. YORK, SILICON VALUES: THE FUTURE OF FREE SPEECH UNDER SURVEILLANCE CAPITALISm (2021); Ben Bradford et al., *Report Of The Facebook Data Transparency Advisory Group*, JUSTICE COLLABORATORY (2019), https://academyhealth.org/sites/default/files/facebookdatatransparencyadvisorygraoupreport52119.pdf; TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

[3] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

key shifts for transforming platform governance, focusing on the structural changes required to both repair and reduce harm.

Position Statement

Researchers are not separate from the social processes they study; our values, beliefs, and experiences inevitably influence our analyses. As such, it is not possible to appropriately position any work without first understanding the relative position of its authors. Both authors of the present work are cisgender women; one author is queer. One author is white, and the other is white-presenting; though we draw from foundational scholarship by a range of scholars to support our analyses, the absence of experiences from or interpretations by Black, Indigenous, and people of color is a significant limitation of this work. It is similarly limited in its cultural perspective, with both authors having lived, been educated, and been employed in the United States. Although one author's experiences of disability inform her perspective, disability justice is also out of scope for the present work. Finally, one author is an academic researcher and tenured professor at a research institution in the midwestern United States; the other is a student at this same institution and has worked as a corporate social media researcher for four years.[4] Both authors are social media users, have personally experienced online harassment, and have studied intersections between social media behavior and governance in both academia and industry.

[4] Blackwell has worked full-time at Facebook and Twitter. Schoenebeck has consulted with Twitter and received funding from Instagram, Facebook, Mozilla, and Google. This work was not directed by, nor does it express the opinions of, any company.

PARADIGMS OF SOCIAL MEDIA GOVERNANCE

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users, including but not limited to flaming (or the use of inflammatory language, name calling, or insults); doxing (or the public release of personally identifiable information, such as a home address or phone number); impersonation (or the use of another person's name or likeness without their consent); and public shaming (or the use of social media sites to humiliate a target or damage their reputation). While online harassment is sometimes depicted as an outlier or fringe behavior, an overwhelming number of social media users have experienced or witnessed some form of online harassment.[5] Harassment tactics are sometimes employed concurrently, particularly when many individuals, acting collectively, target just one individual (sometimes referred to as "dogpiling"). One individual may also harass another, as is often the case in instances of cyberbullying[6] and non-consensual intimate image sharing (also known as "revenge porn"), in which sexually explicit images or videos are distributed without their subject's consent, often by a former romantic partner.[7] Online harassment experiences can range from a single instance to repeated harassment over a sustained period of time; similarly, given the networked

---

[5] Maeve Duggan, *Online Harassment 2017*, PEW RESEARCH CENTER: INTERNET & TECHNOLOGY (Jul. 11, 2017), http://www.pewinternet.org/2017/07/11/online-harassment-2017/.

[6] Zahra Ashktorab & Jessica Vitak, *Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design With Teenagers*, *in* PROCEEDINGS OF THE 2016 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYS. 3895 (2016); Peter K. Smith et al., *Cyberbullying: Its Nature and Impact in Secondary School Pupils*, 49 J. CHILD PSYCH. & PSYCHIATRY 376 (2008).

[7] CARRIE GOLDBERG, NOBODY'S VICTIM: FIGHTING PSYCHOS, STALKERS, PERVS, AND TROLLS (2019); Danielle Keats Citron, *A New Compact for Sexual Privacy*, William & Mary L.R. (forthcoming), https://papers.ssrn.com/abstract=3633336 (last visited Dec. 7, 2020); Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345 (2014).

nature of social media platforms, targets may be harassed by one perpetrator or thousands. These attributes often overlap, especially in the case of coordinated, networked harassment campaigns that are long-term and large-scale.

Regulating behavior is complex, and contemporary social media platforms face numerous challenges. Some are challenges of scale: monolithic approaches to online governance approaches start to crumble at the scale of millions or even billions of diverse users.[8] Others are challenges of adaptability: best practices in one community or platform may fall short in another, particularly on large, global platforms where diverse individual and cultural norms intersect. They may also be failures of anticipation: few could have foreseen the concentration of global power now held by a handful of corporate leaders.

Social media governance is both social and technical; the sociotechnical perspective[9] describes how social and technical aspects of systems are necessarily interrelated and cannot be disentangled. In other words, we cannot design a technological system without also considering its social impacts, and we cannot understand the social impacts of a system without also considering its design and politics. A sociotechnical lens of social media governance argues that design principles and practices will inevitably shape how social behavior is governed online, and vice versa. This section establishes four major paradigms of social media governance: normative, distributed, algorithmic, and retributive

---

[8] GILLESPIE, *supra* note 2; ROBERTS, *supra* note 3.

[9] Mark S. Ackerman, *The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility*, 15 HUM.–COMPUT. INTERACTION 179 (2000).

regulation.[10] These paradigms are overlapping, both temporally and categorically, and reflect evolving social behaviors and technological affordances.

Normative Regulation

The earliest paradigm of governance, emerging in the 1980s[11], involved establishing and reinforcing norms for good behavior, sometimes assigning community members special privileges (e.g., administrator or moderator status) to enforce those norms.[12] This early paradigm also saw the introduction of specialized moderation tools to support regulation, such as reporting, flagging, and editorial rights.[13]

Online communities continue to rely on normative regulation today, both through formal rules—typically asserted by community guidelines and enforced via content moderation[14]—as well as through unstated, informal norms that are learned through

---

[10] An early version of these paradigms was developed in Lindsay Blackwell et al., *When Online Harassment is Perceived to be Justified*, *in* INTERNATIONAL AAA CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM) (2018).

[11] HOWARD RHEINGOLD, THE VIRTUAL COMMUNITY: HOMESTEADING ON THE ELECTRONIC FRONTIER (2000); JULIAN DIBBELL, MY TINY LIFE: CRIME AND PASSION IN A VIRTUAL WORLD (1998).

[12] Eshwar Chandrasekharan et al., *The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales*, 2 PROC. ACM HUM.-COMPUT. INTERACT. 32:1 (2018); DIBBELL, supra note 11; Robert Kraut & et al., *The HomeNet Field Trial of Residential Internet Services*, 39 Commc'n of the ACM 55 (1996); ROBERT E. KRAUT ET AL., BUILDING SUCCESSFUL ONLINE COMMUNITIES: EVIDENCE-BASED SOCIAL DESIGN (2012); Cliff Lampe & Paul Resnick, *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Spac*e, in PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 543 (2004).

[13] Lindsay Blackwell et al., *Classification and its Consequences for Online Harassment: Design Insights from Heart*Mob, 1 PROC. ACM HUM.-COMPUT. INTERACT. 19 (2017); J. Nathan Matias et al., *Reporting, Reviewing, and Responding to Harassment on Twitter* (2015), http://womenactionmedia.org/twitter-report; Jessica A. Pater et al., *Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms*, in PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON SUPPORTING GROUP WORK 369 (2016).

[14] ROBERTS, *supra* note 3.

participation in the community.[15] While social media companies have largely relied on prescriptive norms (i.e., explicit rules) to govern user behavior, descriptive norms—the implicit social expectations we learn by observing how others interact in a given space—are much more powerful at shaping behavior. Prescriptive norms establish how people *should* behave, descriptive norms describe how people are already behaving—creating what Cialdini describes as "a decisional shortcut" when other people are choosing how to behave.[16]

Although normative regulation allows communities to self-govern in ways that are aligned with their specific values and priorities, these strategies are more effective in communities with clearly-established boundaries, such as individual subreddits.[17] Many popular platforms, such as Twitter and TikTok, lack formal community infrastructures, which constrains their ability to rely on normative regulation. Even in online spaces with a clear sense of community, antisocial norms—for example, norms that encourage discrimination, hatred, racism, and other harms—may also emerge and can persist if left unchecked.[18]

Distributed Regulation

A second paradigm saw the rise of crowd-sourced approaches to behavioral regulation, first popularized by platforms

---

[15] J. Nathan Matias, *Preventing Harassment and Increasing Group Participation Through Social Norms in 2,190 Online Science Discussions*, 116 PNAS 9785 (2019); Chandrasekharan et al., supra note 12.
[16] Robert B. Cialdini, Carl A. Kallgren & Raymond R. Reno, *A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior*, 24 ADVANCES IN EXPERIMENTAL SOC. PSYCH. 201 (1991).
[17] Matias, *supra* note 15; Chandrasekharan et al., *supra* note 12.
[18] Eshwar Chandrasekharan et al., *You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech*, 1 PROC. ACM HUM.-COMPUT. INTERACT. 31:1 (2017); Kishonna L. Gray, *Black Gamers' Resistance*, *in* RACE AND MEDIA: CRITICAL APPROACHES 241 (Lori Kido Lopez ed., 2020).

in the early 2000s (e.g., Slashdot and Digg) and still in use by some contemporary platforms (e.g., Reddit and Wikipedia). This model of governance—what Grimmelmann characterizes as distributed moderation[19]—traditionally relies on scalable feedback mechanisms (e.g., upvotes and downvotes) to establish the appropriate enforcement action. For example, a post that receives a high volume of upvotes may be featured more prominently; conversely, a post receiving a high volume of downvotes may be a candidate for deletion.

Distributed and normative regulation overlap in their reliance on shared community norms to govern behavior. Thus, while crowd-sourced governance can be an effective mechanism for reducing harmful content, this is ultimately dependent on the specific values of a given community. Some communities may embrace offensive, violent, or other kinds of damaging content as desirable,[20], rendering distributive regulation effective at enforcing the community's values but not at discouraging harm. Distributed moderation is also vulnerable to manipulation; most technical feedback mechanisms are easily manipulated by smaller factions of users (e.g., recruiting additional users to artificially inflate vote counts), sometimes with the express purpose of amplifying harm.

Algorithmic Regulation

A third paradigm of regulation—and the dominant governance mechanism for large social media companies, such as Facebook, Twitter, and YouTube—relies on automated techniques

---

[19] James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42 (2015); Lampe & Resnick, *supra* note 12.

[20] Michael Bernstein et al., *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*, *in* INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM) 50 (2011).

for evaluating potentially harmful content.[21] This class of strategies uses machine learning and natural language processing to develop computational models that systematically evaluate large quantities of data.

To facilitate scaled content moderation, machine learning models are typically trained to detect language that may be abusive or violent,[22] often automatically removing entities at a certain level of model confidence. Although automated content moderation approaches continue to improve, accurate and reliable detection is challenging at best, even in far less complex applications than the detection of nuanced behaviors like online harassment and hate speech. Social media companies have to make necessary trade-offs between a model's precision (i.e., accuracy) and its recall, or the quantity of relevant instances the model returns. They often optimize for recall out of necessity—nearly a billion tweets are sent per day[23]—resulting in imprecise models plagued by false positives, where harmful content evades detection (where permissible content is incorrectly removed), and true negatives (where harmful content evades detection).

Contrary to popular perception, algorithmic regulation does not eradicate the need for human input. Supervised learning

---

[21] This has been referred to as the "industrial approach" in Robyn Caplan, *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches,* DATA & SOCIETY (2018).

[22] Eshwar Chandrasekharan et al., *The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data*, *in* PROCEEDINGS OF THE 2017 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 3175 (2017); Hossein Hosseini et al., *Deceiving Google's Perspective API Built for Detecting Toxic Comments* (2017), http://arxiv.org/abs/1702.08138; Ellery Wulczyn, Nithum Thain & Lucas Dixon*, Ex Machina: Personal Attacks Seen at Scale*, in PROCEEDINGS OF THE 26TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 1391 (2017); Dawei Yin et al., *Detection of Harassment on Web 2.0*, *in* PROCEEDINGS OF THE CONTENT ANALYSIS IN THE WEB 2.0 WORKSHOP (2009).

[23] Twitter Usage Statistics, https://www.internetlivestats.com/twitter-statistics/ (last visited May 31, 2021).

models—i.e., a machine learning model that predicts the similarity between a given piece of text and the dataset used to "teach," or train, the model—requires high volumes of annotated data, typically labeled by humans, both for training initial models and for refining their performance over time. Although investing in algorithmic regulation will relieve some burden from workers—companies with well-performing algorithms can, over time, rely on fewer workers for manual content moderation—machine learning still requires a sizeable workforce of human laborers to review hateful, violent, and otherwise traumatizing content over long shifts and for low wages.[24]

Finally, automated governance is also relatively easy to bypass through subtle modifications of language.[25] When combined, these limitations can result in harmful content persisting on social media while jokes, cultural references, and in-group conversations are, from the user's perspective, inexplicably removed.

Retributive Regulation

A fourth paradigm of governance, which has risen to prominence most recently, reflects a complex spectrum of conditions in which social media users aspire to enforce justice themselves—in part due to the recognized failures of social media companies to adequately govern their platforms.[26] When offenders are not held accountable for their actions, users may instead turn to moral shaming to enact retribution[27]—resulting in punishments that, as Kate Klonick argues, may be indeterminate, uncalibrated, or inaccurate.

---

[24] ROBERTS, *supra* note 3.
[25] Hossein Hosseini et al., *supra* note 22.
[26] Lindsay Blackwell et al., *Classification and Its Consequences for Online Harassment: Design Insights from HeartMob*, 1 PROCS. OF THE ACM ON HUM.-COMPUT. INTERACTION (2017).
[27] JON RONSON, SO YOU'VE BEEN PUBLICLY SHAMED (2016).

An individual user leveraging social media to retaliate against a perceived offender may seem unremarkable; however, the affordances of networked platforms can escalate ordinary social sanctioning into something resembling mass vigilantism. Social feedback (such as likes or upvotes) and algorithmic amplification promote perceptions of endorsement that can result in large-scale group behaviors, which often have extreme and disproportionate impacts on perceived offenders—including threats to physical safety, unwanted disclosures of personal information, sustained social isolation, and job loss.[28]

Retributive regulation is sometimes crudely collapsed into a single set of behaviors, without consideration for the kinds of injustices or harms that necessitate those behaviors. For example, so-called "cancel culture"—a neologism describing a type of mass social sanctioning in which a person's social or professional status is questioned due to a perceived infraction—has arisen as one outcrop of this fourth governance paradigm. Characterizations about the existence of cancel culture should be evaluated cautiously; Meredith Clark argues that the label is often misused, with justifiably critical responses to legitimately harmful acts regularly dismissed as "cancel culture" without recognition of the desired accountability.[29]

This most recent paradigm shift, coupled with the proliferation of online misinformation and increasing political discord, has accelerated demands for formal regulation to hold social media companies accountable for the ramifications of inadequate platform governance. These demands coincide with

---

[28] RONSON, *supra* note 27; GOLDBERG, *supra* note 7; Citron, *A New Compact for Sexual Privacy*, *supra* note 7.
[29] Meredith D. Clark, *DRAG THEM: A Brief Etymology of So-Called "Cancel Culture"*, 5 COMMC'N & PUB. 88 (2020).

ongoing discussions about the possibilities and limitations for users and communities to regulate themselves.[30]

HARMS DUE TO INADEQUATE SOCIAL MEDIA GOVERNANCE

Harm refers to damage, injury, or setbacks toward a person, entity, or society. Some harms are small and easily repairable, such as the theft of a bicycle. Others, such as the loss of health, are irreparable and cannot be adequately compensated. Harm is distinct from violence, though they are linked; violence will by definition typically cause harm. Harm is a complex and varied concept without a single definition or interpretation; what constitutes harm will vary with use and context. In legal contexts, harm refers to loss or damage to a person's right, property, or well-being, whether physical or mental. In Internet law, scholars have argued for legal recognition of particular kinds of privacy harms,[31] data breach harms,[32] and intimate data harms.[33] Our focus lies in sociotechnical harms—the online content or activity that inflicts psychological or psychological damage towards a person or community and that compromises their ability to participate safely and equitably both online and offline."

Social media platforms facilitate myriad harms, from sexual harassment to hate speech to racism to disinformation. These harms can be intentional (e.g., doxxing a journalist because she wrote something somebody did not like) or unintentional (e.g., sharing content on Twitter that may be inaccessible to disabled people). Intent is a slippery concept to measure; someone intending to be helpful or supportive may still cause harm regardless, in the same

---

[30] Joseph Seering, *Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation*, 4 PROC. ACM HUM.-COMPUT. INTERACT. 107:1 (2020).
[31] Ryan Calo, *The Boundaries of Privacy Harm Essay*, 86 IND. L.J. 1131 (2011).
[32] Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 TEX. L. REV. 737 (2017).
[33] Citron, *A New Compact for Sexual Privacy*, *supra* note 7.

way that someone who intends to cause harm may claim otherwise when facing undesirable consequences. Additionally, harmful experiences can be differentially traumatic to different people and groups.

We consider two predominant, intersecting categories of harms: platform-perpetrated harms (i.e., those perpetrated by the design of platforms) and platform-enabled harms (i.e., those facilitated by platforms but perpetrated by users or groups). These categories build on our stance that consequences of inadequate platform governance are the responsibility of the platforms themselves.

Psychological Distress

Interpersonal abuse, such as online harassment and hate speech, is widespread and can be profoundly damaging for both targets and bystanders. The effects of harassment vary from person to person, ranging from anxiety, humiliation, and self-blame to anger and physical illness.[34] Online harassment in particular can "cast a long shadow," due in part to the persistence and searchability of digital media—severe harassment can inflict long-term damage to an individual's reputation, comfort, or safety. Perhaps most critically, online harassment has a chilling effect on future disclosures: Lenhart et al. found that, in 2016, 27% of American internet users were self-censoring what they post online due to fear of harassment.[35]

Thus, although harassment is instantiated online, targets of online harassment frequently report disruptions to their offline lives,

---

[34] Maeve Duggan, *Online Harassment*, PEW RESEARCH CENTER (Oct. 22, 2014), https://www.pewresearch.org/internet/2014/10/22/online-harassment/.
[35] Amanda Lenhart et al., *Online Harassment, Digital Abuse, and Cyberstalking in America*, DATA & SOCIETY (2016), https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/.

including emotional and physical distress, changes to technology use or privacy behaviors, and increased safety and privacy concerns. People who experience harassment often choose to temporarily or permanently abstain from social media sites, despite the resulting isolation from information resources and support networks. Online harassment can also be disruptive to personal responsibilities, work obligations, and sleep due to the labor of reporting harassment to social media platforms or monitoring accounts for activity. Some types of online harassment specifically aim to disrupt a target's offline life, such as swatting (i.e., falsely reporting a crime to encourage law enforcement agencies to investigate a target's home or business).

Online abuse can also result in fear for one's physical safety, regardless of whether or not threats of physical harm ever materialize. Revealing a person's home address, for example, results in a loss of perceived security that endures even after any online harassment has ceased[36]—highlighting the tangible impact of even a potential for harm on the ability for social media users to live safely and comfortably.

Physical Violence

Numerous studies demonstrate the correlation between inciting language online and subsequent offline violence, particularly when social media is used to stoke existing physical conflict. Desmond Patton and coauthors have described the use of social media by gang-involved youth to levy taunts and threats against rival groups, often in response to romantic conflict or expressions of grief and amplified by the affordances of social

[36] *See* stories in GOLDBERG, *supra* note 7.

media platforms.[37] The rapid exchange of comments, pictures, and videos between existing rivals—exacerbated by the network-based visibility of social media content[38]—intensifies any perceived slights, increasing the likelihood of online conflict escalating to physical fights. This perpetuates a cycle of physical and emotional violence in which young people struggling with loss turn to social media for support and instead find themselves embroiled in additional conflict.[39]

Facebook has acknowledged its platform's role in fomenting ethnic violence in Myanmar, in large part due to the deliberate spread of misinformation used to stoke pre-existing tensions between Myanmar's majority-Buddhist population and the Rohingya, a minority Muslim community subjected to ongoing persecution by military and state actors.[40] Despite warnings by researchers and human rights activists about the proliferation of Burmese hate speech on its platform, investigative journalists continued to find hate speech, threats of violence, and calls for genocide on the platform.[41] Similarly, Twitter itself has recognized its role in the January 6, 2021 "storming" of the US Capitol building which resulted in violence, destruction, and fatalities. Soon after the

---

[37] Desmond Upton Patton et al., *Internet Banging: New Trends in Social Media, Gang Violence, Masculinity and Hip Hop*, 29 COMPUT. IN HUM. BEHAV. A54 (2013); Desmond Upton Patton et al., *You Set Me Up: Gendered Perceptions of Twitter Communication Among Black Chicago Youth*, 6 SOC. MEDIA & SOCIETY (2020); Desmond Upton Patton et al., *Expressions of Loss Predict Aggressive Comments on Twitter Among Gang-Involved Youth in Chicago*, 1 NPJ DIGITAL MEDICINE 1–2 (2018).

[38] Caitlin Elsaesser et al., *Small Becomes Big, Fast: Adolescent Perceptions of How Social Media Features Escalate Online Conflict to Offline Violence*, 122 CHILD. & YOUTH SERVICES REV. 122 (2021).

[39] Patton et al., *Internet Banging*, *supra* note 38.

[40] Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, THE NEW YORK TIMES, (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

[41] Steve Stecklow, *Why Facebook Is Losing The War on Hate Speech in Myanmar*, REUTERS (Aug. 15, 2018), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/; Stevenson, *supra* note 41.

insurrection, and after repeated calls for the removal of inciting tweets by then-President Donald Trump, Twitter permanently removed Trump's account, citing risks of further violence.[42]

Similar violence around the world has been associated with the proliferation of misinformation and hate speech on social media platforms. The circulation of rumors on WhatsApp—an encrypted chat client owned by Facebook—has contributed to a rise in mob lynchings across India.[43] In post-war Sri Lanka, increased violence against Muslim communities and other religious minorities has coincided with an increase in the country's social media users, particularly among Sinhalese Buddhists. [44] In the United States, numerous acts of white supremacist violence were perpetrated by domestic extremists who participated in radical online forums (e.g., Gab, Parler, 4chan).[45] In Pakistan, women are have been silenced through threats of, or actual, violence and death; in 2016, ongoing harassment of Qandeel Baloch, a social media celebrity and activist, culminated in her murder by her own brother.[46]

While threats of physical violence can be delivered on any social media user or community, they often reflect existing disparities between populations: those who are able to exist safely in their homes and local communities may also be able to be safer

---

[42] *Permanent suspension of @realDonaldTrump*, TWITTER (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html.
[43] Chinmayi Arun, *On WhatsApp, Rumours, Lynchings, and the Indian Government*, 54 ECON. & POL. WKLY. (2019).
[44] Sanjana Hattotuwa, *Digital Blooms: Social Media and Violence in Sri Lanka,* TODA PEACE INSTITUTE, 12 (2018), https://toda.org/assets/files/resources/policy-briefs/t-pb-28_sanjana-hattotuwa_digital-blooms-social-media-and-violence-in-sri-lanka.pdf.
[45] Laurel Wamsley, *On Far-Right Websites, Plans To Storm Capitol Were Made In Plain Sight*, NPR (Jan. 7, 2021), https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/07/954671745/on-far-right-websites-plans-to-storm-capitol-were-made-in-plain-sight.
[46] Imran Gabol & Taser Subhani, *Qandeel Baloch murdered by brother in Multan: police*, DAWN (July 16, 2016), http://www.dawn.com/news/1271213.

online, while those who experience discrimination and persecution offline may be similarly vulnerable online.

Oppression and Marginalization

We cannot talk about harm without also talking about power, because power differences are structural enablers of harm. Power enables abuse through its facilitation of transgressions and its dismantling of accountability. Power differentials manifest in interpersonal contexts (e.g., based on gendered hierarchies)[47] as well as in organizational contexts (e.g., based on workplace hierarchies).[48] Power differentials also arise in emergent ways on social media; influencer status and follower counts provision enormous power to users who gain those statuses or counts,[49] without guidance for or calibration around wielding that power appropriately. Around the world, vulnerable social media users, including dissidents, women, people of color, refugees, transgender people, and members of other non-dominant social groups experience disproportionate harm in online contexts.[50] These experiences are often overlooked, dismissed, or exacerbated by systems of platform governance that fail to account for or even acknowledge the systemic power disparities that enable them.

Technology reflects—and often exacerbates—structural inequities that persist in society writ large. While platforms bear

---

[47] Christopher Uggen & Amy Blackstone, *Sexual Harassment as a Gendered Expression of Power*, 69 AM. SOCIO. REV. 64 (2004).
[48] *Id*.
[49] TAMA LEAVER, TIM HIGHFIELD & CRYSTAL ABIDIN, INSTAGRAM: VISUAL SOCIAL MEDIA CULTURES (2020).
[50] YORK, *supra* note 2; *Online violence: Just because it's virtual doesn't make it any less real*, GLOBAL FUND FOR WOMEN (2015), https://www.globalfundforwomen.org/online-violence-just-because-its-virtual-doesnt-make-it-any-less-real/; *Toxic Twitter – A Toxic Place for Women*, AMNESTY INT'L (2018), https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.

responsibility for hosting and facilitating harassment, violence, and extremism, these are enduring social problems that cannot be rooted out by social media reform alone. For decades, scholars have documented how racist behavior online intersects with existing offline racism.[51] In 2009, early facial recognition technology developed by HP could easily track the movements of a white user, but failed to recognize black users; later, in 2015, Google's own facial recognition technology categorized pictures of black people as containing images of gorillas.[52] In 2017, despite Apple's efforts to train its own Face ID technology on a large and diverse set of faces,[53] a Chinese woman discovered that her colleague—also a Chinese woman—was able to unlock her device on every attempt.[54] In her book Algorithms of Oppression, Safiya Noble (2018) details countless examples of racial biases that have been "baked in" to the technological systems we use every day: for example, Google returning pictures of white women when queried for images of "professional women," but pictures of black women when queried for images of "unprofessional hair."[55]

Gender and sexual discrimination is also prevalent in technology design, from default avatars registering as male

---

[51] LISA NAKAMURA, CYBERTYPES: RACE, ETHNICITY, AND IDENTITY ON THE INTERNET (2002); JESSE DANIELS, CYBER RACISM: WHITE SUPREMACY ONLINE AND THE NEW ATTACK ON CIVIL RIGHTS (2009); Gray, *supra* note 18.

[52] Klint Finley, *Can Apple's iPhone X Beat Facial Recognition's Bias Problem?*, WIRED (Sept. 13, 2017), https://www.wired.com/story/can-apples-iphone-x-beat-facial-recognitions-bias-problem/.

[53] Kate Conger, *How Apple Says It Prevented Face ID From Being Racist*, GIZMODO (Oct. 16, 2017), https://gizmodo.com/how-apple-says-it-prevented-face-id-from-being-racist-1819557448.

[54] Christina Zhao, *Is the iPhone X's facial recognition racist?*, NEWSWEEK (Dec. 18, 2017), https://www.newsweek.com/iphone-x-racist-apple-refunds-device-cant-tell-chinese-people-apart-woman-751263.

[55] SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

silhouettes[56] to Facebook's ongoing challenges surrounding its "real name" policy and the deactivation of accounts belonging to trans users, drag queens, Indigenous people, abuse survivors, and others whose identities or account names may be inconsistent with their legal names.[57] Most online forms requiring gender information only offer a binary choice—"male" or "female"—forcing non-binary individuals to either choose an incorrect gender category or refrain from using the site or service.[58] The implicit biases designed into everyday technologies not only reflect existing discrimination, but may also exacerbate it: exposure to negative stereotypes about one's social identity can actually reduce performance on a relevant task, a phenomenon known as stereotype threat.[59] Further, these technological biases, however unintentional, are often only identified—and subsequently given the opportunity for correction—through proactive auditing by researchers, in a practice Sandvig, et al. (2014) call algorithmic auditing.[60]

These challenges are partly, though not entirely, due to problems of classification. Social media platforms rely on numerous

---

[56] April H. Bailey & Marianne LaFrance, *Anonymously Male: Social Media Avatar Icons Are Implicitly Male and Resistant to Change*, 10 CYBERPSYCHOLOGY: J. PSYCH. RSCH. ON CYBERSPACE (2016).

[57] Vauhini Vara, *Drag Queens Versus Facebook's Real-Names Policy*, THE NEW YORKER (Oct. 2, 2014), https://www.newyorker.com/business/currency/whos-real-enough-facebook; Oliver L. Haimson & Anna Lauren Hoffmann, *Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities*, 21 FIRST MONDAY (2016).

[58] Scheuerman, Morgan Klaus et al., *Revisiting Gendered Web Forms: An Evaluation of Gender Inputs with (Non-) Binary People*, *in* PROCEEDINGS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (May 2021).

[59] Claude M. Steele, Steven J. Spencer & Joshua Aronson, *Contending with Group Image: The Psychology of Stereotype and Social Identity Threat*, 14 ADVANCES IN EXPERIMENTAL AND SOC. PSYCH. 379 (2002).

[60] Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, in* DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY (2014).

classification systems and categorization schema[61]: algorithmic feeds serve specific content based on particular features; reporting flows ask users to identify specific policy violations; profile creation requires various selections from predefined lists. But when classification systems are built to optimize for scale, variation is flattened in favor of majority experiences. This results in compounding harms for users and communities who are already socially, economically, or otherwise excluded from society. For example, when sex trafficking is prohibited on mainstream platforms, consensual sex work is often caught up in the same algorithmic net; this has the immediate material effect of reduced income for sex workers (who themselves often possess multiple stigmatized identities such as being queer or non-white), while also contributing to the continued stigmatization of sex-based labor.[62] The embedded biases inherent in large-scale automation manifest in many forms, across gender, race, disability, and other characteristics—most acutely at their intersections—and often in ways that are not transparent or interpretable to the users whose experiences are governed by them.

Threats to Free Expression

Regulatory recommendations typically focus on refinements to specific legislation. In the U.S., scholars have called for "reasonable moderation practices rather than the free pass" that is enabled by 47 U.S.C. § 230, a provision of the Communications Decency Act (CDA) of 1996 protecting online service providers from incurring legal liability for third-party (i.e., user-generated)

---

[61] Blackwell et al., *Classification and its Consequences for Online Harassment*, *supra* note 13.

[62] See stories from sex workers documented in Kendra Albert et al., *FOSTA in Legal Context* (2020), https://papers.ssrn.com/abstract=3663898; YORK, *supra* note 2.

content.[63] Platforms frequently cite freedom of expression when deciding to minimize their role in arbitration, a stance buttressed by the "safe harbor" offered by Section 230.[64]

Though Section 230 has had an outsized influence on US-based corporate governance, many regions around the world are debating regulatory practices, with varying thresholds for the types of content social media companies are legally required to remove. In Germany, NetzDG requires platforms to promptly remove illegal content in Germany, including Anti-Semitic speech and hate speech based on religion or ethnicity.[65] In Korea, Article 44 of the Information and Communications Network Act (ICNA) encourages proactive removal of content if requested.[66] In India, the IT Act provides immunity for platforms as long as they take action to address certain categories of content within a short time frame.[67] In Australia, platforms have to moderate and also report "abhorrent violent" content.[68] In other countries, such as Syria, Turkey, Pakistan, and Tunisia, partial or wholesale bans on social media result in widespread censorship of expression by state actors.[69]

---

[63] DANIELLE KEATS CITRON & MARY ANNE FRANKS, *The Internet As a Speech Machine and Other Myths Confounding Section 230 Reform*, U. CHI. L. FORUM (forthcoming, 2020).
[64] *Id*.
[65] *Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information*, BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ [FEDERAL MINISTRY OF JUSTICE AND CONSUMER PROTECTION], https://www.BMJV.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html (last visited Apr 8, 2021).
[66] *Act on Promotion of Information and Communications Network Utilization and Information Protection, etc.*, KOREAN LAW TRANSLATION CENTER, https://elaw.klri.re.kr/eng_service/lawView.do?hseq=38422&lang=ENG (last visited Apr 8, 2021).
[67] The Information Technology Act, 2000 (India).
[68] Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act, 2019 (Austl.).
[69] For a comprehensive discussion of platforms, free speech and censorship, and state governance, see YORK, *supra* note 2.

Freedom of expression is a human right; however, its contours are nuanced and vary by regions and contexts (e.g., attitudes towards nudity, which is considered normative in some cultures but highly sensitive in others). Preserving freedom of expression while also mitigating harm is a complex endeavor. For example, in her book, *Silicon Values*, Jillian York highlights how platforms' automated removal of violent extremist content prompted human rights groups to begin preserving that content as evidence of war crimes.[70] Chinmayi Arun notes that mounting pressure on social media companies to cooperate with governments has alarming implications—both for individual user privacy and the continued utility of these platforms for journalists, activists, and political dissidents.[71] While this article is not focused on the specific nuances of free expression, any proposal for shifts in social media governance must also consider implications for human rights, including the potential for exploitation by state actors.

PRINCIPLES FOR SOCIAL MEDIA GOVERNANCE

Although social media governance to date has largely been informed by Western models of criminal justice, which rely on sanctions (e.g., punishment) to encourage compliance with formal rules and laws, we argue for systems of governance that instead focus on accountability for and repair of specific harms. Social media governance should be informed by both punitive and restorative frameworks; here, we propose how theories of justice can inform social media policies, practices, and products that acknowledge and attend to harm.

---

[70] *Id*.
[71] Chinmayi Arun, *Facebook's Faces*, 135 HARV. L. REV. F. (forthcoming).

Retribution and Punishment

The concept of justice is invoked when deciding how society should respond to a person who is perceived to have committed some infraction (i.e., a violation of rules and laws). In Western societies, criminal justice approaches have traditionally sought to discourage offenders through the fear of strict criminal sanctions. The concept of retribution is focused on delivering offenders their deservedness,[72], and proportionality[73] in criminal sentencing. Moral judgment plays a powerful role in retribution and shapes cultural attitudes, policy, and law around appropriate punishments.[74] Feelings of moral anger and disgust (e.g., feelings that results if someone engages in pedophilia) often protect and preserve social order within a society.[75] In the United States, incarceration has been a predominant engine for enacting punishment, especially towards some groups including people of color, disabled people, and poor people.[76]

Social media governance has typically adopted Western frameworks of criminal justice: identifying perpetrators of undesirable behavior and administering punitive responses.[77] If

---

[72] Kevin M. Carlsmith & John M. Darley, *Psychological Aspects of Retributive Justice*, 40 ADVANCES IN EXPERIMENTAL SOC. PSYCH. 193 (2008); IMMANUEL KANT & WERNER PLUHAR, CRITIQUE OF JUDGMENT (1987).
[73] Michael Wenzel et al., *Retributive and Restorative Justice*, 32 L. HUM. BEHAV. 375 (2008).
[74] Roger Giner-Sorolla et al., *Emotions in Sexual Morality: Testing the Separate Elicitors of Anger and Disgust* 26 COGNITION & EMOTION 1208 (2012); Jesse Prinz, *Is Morality Innate?*, *in* MORAL PSYCHOLOGY: THE EVOLUTION OF MORALITY: ADAPTATIONS AND INNATENESS 608 (Walter Sinnott-Armstrong & Christian B. Miller eds., 2007).
[75] Bunmi O. Olatunji & Craig N. Sawchuk, *Disgust: Characteristic Features, Social Manifestations, and Clinical Implications*, 24 J. SOC. & CLINICAL PSYCH. 932 (2005); Pascale Sophie Russell & Roger Giner-Sorolla, *Moral Anger, but Not Moral Disgust, Responds to Intentionality*, 11 EMOTION 233 (2011).
[76] RUEBEN JONATHAN MILLER, HALFWAY HOME (2021).
[77] Bradford et al., *supra* note 2; Eshwar Chandrasekharan et al., *supra* note 2; Shagun Jhaver, Amy Bruckman & Eric Gilbert, *Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations*

content is found to violate a platform's community guidelines, platform responses range from removing the content or demoting its visibility to banning the user who produced it, either temporarily or permanently. However, these sanctions embrace many of the problems of retributive models of governance; namely, they overlook the needs and interests of the targets of harassment and remove offenses and offenders from the community without any attempt at rehabilitation. Contemporary platform governance also relies on obfuscated processes of content moderation that have little transparency or accountability to all involved parties[78]; content is deleted without leaving any visible trace of its removal; policy violators have little opportunity for recourse and may not even be informed of the specific rule they have broken; reporters receive generalized responses that often don't reference the content in question, if they receive a response at all.

In typical platform-driven moderation systems, all violators are treated equally, with users who unintentionally violate rules receiving the same sanctions as users who deliberately try to cause harm. Instead, we argue for an expanded set of remedies, one that better recognizes and remediates harms by incorporating responsive penalties that allow for reeducation, rehabilitation, and forgiveness. Social media users already intuitively imagine diverse and varying punishments that allow for proportional responses to varied infractions, depending both on the specific type of violation and the perceived intent of the violator.[79] For example, people who

---

on Reddit, CSCW PROC. ACM HUM.-COMPUT. INTERACT. (2019); J. Nathan Matias, *supra* note 15; Pater et al., *supra* note 13; Sarah Perez, *Twitter adds more anti-abuse measures focused on banning accounts, silencing bullying*, TECHCRUNCH (Mar. 1, 2017), http://social.techcrunch.com/2017/03/01/twitter-adds-more-anti-abuse-measures-focused-on-banning-accounts-silencing-bullying/.

[78] The Santa Clara Principles, *supra* note 2.

[79] Lindsay Blackwell et al., *Harassment in Social Virtual Reality: Challenges for Platform Governance*, 3 PROC. ACM HUM.-COMPUT. INTERACT. 100:1 (2019).

perpetuate one-time or occasional offenses can be given the opportunity to correct and make amends for their behavior, with more severe penalties reserved for users who perpetuate sustained abuse without remorse.

Moderation practices that eschew blunt, one-size-fits-all penalties in favor of sanctions which are proportionate to specific violations is aligned with what Braithwaite calls responsive regulation.[80] In a responsive regulation framework, the least interventionist punishments—for example, education around existing rules and policies—are applied to first-time or other potentially redeemable offenders, with sanctions for repeat violators escalating in severity until they reach total incapacitation (e.g., a permanent account- or IP address-level ban).[81] By implementing enforcement decisions that are responsive to the context of specific infractions, platforms may be perceived as more legitimate when harsher penalties are required: a user won't become eligible for permanent suspension without being given multiple opportunities to correct their behavior and adhere to platform policies. Responsive regulation may also help platforms avoid alienating users for incorrect enforcement decisions; when the full context surrounding a violation is unclear, a less severe penalty can be applied.

Accountability and Restoration

Alternative justice models for platform governance could recognize harm, establish accountability for that harm, and establish an obligation to repair harm. Whereas a retributive justice governance approach would ask what laws have been broken, who broke them, and what punishment is deserved, alternative justice

---

[80] IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION: TRANSCENDING THE DEREGULATION DEBATE (1992).

[81] JOHN BRAITHWAITE, RESTORATIVE JUSTICE & RESPONSIVE REGULATION (2002).

approaches would instead ask who has been harmed, what do they need, and how should systems be redesigned to prevent harms from reoccurring? However, alternative justice systems are not in themselves sufficient to address harm; any justice system that is implemented—whether traditional or alternative—may inadvertently protect and benefit social groups who are already privileged unless the systems are explicitly designed to do otherwise.

Two prominent alternative justice frameworks are restorative justice and transformative justice. Restorative justice is a framework and movement that encourages mediated conversations between those who perpetuate and those who experience harm, typically with mediators and community members actively participating. Restorative justice asks that offenders acknowledge wrongdoing and harm, accept responsibility for their actions, and express remorse. Restorative justice has been practiced in Indigenous communities, and has been advanced as an alternative to Western criminal justice systems that over-incarcerated Indigenous youth. In New Zealand, for example, restorative justice was the foundation for a 1989 act between Maori people and New Zealand Parliament which was designed to care for Indigenous children rather than moving them into prison pipelines.[82]

Recognition of wrongdoing is an essential first step in establishing accountability for harm. The concept of recognition is often invoked in human rights discussions and contains two facets: recognition of human rights, and recognition of violations of those rights. However, recognition has also been misused as a politicized form of collective identity that demands recognition of a dominant group while perpetuating distributive injustices towards non-

[82] The Oranga Tamariki Act, 1989 (N.Z.).

dominant groups.[83] Restorative justice programs were sometimes implemented without consideration of race or disability;[84]; as a result, able bodied white women offenders might have been viewed as victims of circumstance who deserved empathy, while disabled people of color continued to be over-incarcerated.[85] Many restorative justice practitioners have chosen to work outside of criminal legal systems because of the ongoing failures of those systems. Thus, recognition is not simply a decision to acknowledge harms, but a confluence of decisions about what rights people should have, how to acknowledge those rights, and how to acknowledge violations of those rights.

Recognition of harm on social media asks for recognition of the multitudes of ways that users and communities can experience harms, including those that fall outside of current regulatory capture. Accountability, then, requires accepting responsibility for those harms, including the obligation to repair them. Scholars Mia Mingus and Mariame Kaba have argued for moving away from holding others accountable and towards supporting proactive accountability, i.e., "active accountability."[86] Centering accountability and repair

---

[83] Nancy Fraser, *Rethinking Recognition: Overcoming Displacement and Reification in Culture Politics*, *in* RECOGNITION STRUGGLES AND SOCIAL MOVEMENTS: CONTESTED IDENTITIES, AGENCY AND POWER (2003).

[84] Theo Gavrielides, *Bringing Race Relations Into the Restorative Justice Debate: An Alternative and Personalized Vision of "the Other"*, 45 J. BLACK STUD. 216 (2014).

[85] Danielle Dirks et al., *'She's White and She's Hot, So She Can't Be Guilty': Female Criminality, Penal Spectatorship, and White Protectionism*, 18 CONTEMP. JUST. REV. 160 (2015).

[86] Mariame Kaba et al., *When It Comes to Abolition, Accountability Is a Gift*, BITCH MEDIA , https://www.bitchmedia.org/article/mariame-kaba-josie-duffy-rice-rethinking-accountability-abolition (last visited Jan 6, 2021); Mariame Kaba & John Duda, *Towards the Horizon of Abolition: A Conversation With Mariame Kaba* (2018), https://transformharm.org/towards-the-horizon-of-abolition-a-conversation-with-mariame-kaba/ (last visited Jan 8, 2021); Mia Mingus, *The Four Parts of Accountability: How To Give A Genuine Apology Part 1*, LEAVING EVIDENCE (Dec. 18, 2019), https://leavingevidence.wordpress.com/2019/12/18/how-to-give-a-good-apology-part-1-the-four-parts-of-accountability/.

requires shifts towards the needs of those harmed, and accountability from those who perpetuate harm. Acts like apologies, mediated conversation, proclamations, and commemorations could all be supported in online interactions as non-material forms of restoration and accountability.[87] For example, apologies can be powerful illocutionary devices for amending wrongdoings, though they need to be genuine or they can further magnify harm, especially for groups who have already experienced oppression.[88] Similarly, intent not to commit harm again, and subsequent actions, can be a form of accountability and restoration. These boundaries could be built into the design of social media sites where targets of harassment could be granted agency to decide whether to engage further, and if so, under what terms. Other acts like compensation or amplification could enact material remedies, which may be important for correcting some kinds of online injustices. While accountability processes hopefully result in resolution, that may not always be attainable, and the burden of reaching resolution should not fall on those who have experienced harm.[89]

Transformative justice, which extends restorative principles and practices beyond individual reconciliation and towards

---

[87] Our prior studies show that U.S. adults and young adults are generally favorable towards the idea of apologies after online harassment. *See* Sarita Schoenebeck, et al., *Drawing from Justice Theories to Support Targets of Online Harassment*, 23 NEW MEDIA & SOCIETY 1278 (2020); Sarita Schoenebeck et al., *Youth Trust in Social Media Companies' Responses to Online Harassment*, PACM HUM.-COMPUT. INTERACTION 2:1 (2021).

[88] Schoenebeck et al., Drawing from Justice Theories to Support Targets of Online Harassment, supra note 88; Schoenebeck et al., *Youth Trust in Social Media Companies' Responses to Online Harassment*, *supra* note 88. While apologies can be a conduit for justice, the delivery of an apology should not create an expectation of forgiveness from the target, nor should it imply that accountability was present.

[89] John Braithwaite, *Restorative Justice: Assessing Optimistic and Pessimistic Accounts*, 25 CRIME & JUSTICE 1 (1999); Jung Jin Choi, Gordon Bazemore & Michael J. Gilbert, *Review of Research on Victims' Experiences in Restorative Justice: Implications for Youth Justice*, 34 CHILD. & YOUTH SERVS. REV. 35 (2012).

systematic change, has been similarly developed and advanced by non-dominant social groups, including immigrant, Indigenous, Black, disabled, and queer and trans communities.[90] Transformative justice involves practices and politics focused on ending sexual violence using processes outside of carceral policing systems. Transformative justice movements propose that prison and state systems create more harm, violence, and abuse rather than addressing them. Two tenets are that violence and abuse should be responded to within communities rather than by criminal legal systems (while noting that communities themselves can also perpetuate violence), and that any responses should combat, rather than reinforce, oppressive societal norms. Transformative justice movements seek not only to respond to current violence, but to address cycles of violence by transforming the conditions that allowed it to happen.

While restorative justice and transformative justice are distinct movements with different principles, they share a commitment to recognizing harm and violence and resisting the carceral systems that perpetuate them. These commitments help to shed light on the failures of current platform governance practices; when platforms fail to explicitly acknowledge and combat existing inequity, they further entrench those harms with content moderation policies that may seem appropriate on an individual level (e.g., disallowing hate speech), but which obscure and perpetuate violence at a structural level (e.g., equating hate speech against men with that against women, which overlooks gender-based oppression). Many

---

[90] BEYOND SURVIVAL: STRATEGIES AND STORIES FROM THE TRANSFORMATIVE JUSTICE MOVEMENT (2020); Sara Kershnar et al., *Toward Transformative Justice*, GENERATION FIVE (2007), http://www.usprisonculture.com/blog/wp-content/uploads/2012/03/G5_Toward_Transformative_Justice.pdf; Mia Mingus, *Transformative Justice: A Brief Description*, LEAVING EVIDENCE (Jan. 9, 2019), https://leavingevidence.wordpress.com/2019/01/09/transformative-justice-a-brief-description/.

approaches to platform governance can be characterized as "reformist reforms"[91] a term for reforms which maintain the status quo by upholding existing oppression systems. In policing, non-reformist reforms would include those that reduce, rather than maintain, the power by police themselves; reformist reforms would be those which instead increase police funding (e.g., body cameras) or scale (e.g., community policing), effectively maintaining or even strengthening the existing systems. Content moderation discussions can easily fall into reformist reform traps—they tweak, tune, and slightly improve what content is moderated and how, while cementing in place governance structures that continue to overlook harms.

Repairing harms is not one-size-fits-all, however; different harms may be paired with different frameworks and approaches, and multiple approaches could be combined together.[92] Any design-centered approach must be recognizant of its own limitations; much as a school cannot overcome economic inequality or a prison cannot overcome racism, design cannot repair the underlying systemic injustices it facilitates. Instead, like restorative and transformative justice movements in schools and prisons, design as a praxis should aim to acknowledge and mitigate harms within those sites, while also questioning the underlying systems that enable those harms. Any system of justice—whether traditional or alternative—may inadvertently protect and benefit social groups who are already privileged unless they are explicitly designed to do otherwise.

---

[91] Kaba & Duda, *supra* note 87.

[92] Eric Goldman, *Content Moderation Remedies*, MICHIGAN TECH. L. REV. (forthcoming), https://papers.ssrn.com/abstract=3810580 (last visited Mar 31, 2021).

Principles for Repairing Harms

We propose several key shifts for social media companies to facilitate the design and development of platform governance models centered on the recognition and repair of harm.

From Neutral to Principled

Social media companies have typically adopted a "neutral" stance, embracing a veneer of impartiality that ostensibly serves to absolve them of the responsibility to adjudicate harm. This aspirational objectivity may be buttressed by an orientation toward measurement, labels, classification, and formalization in how technology is produced.[93] Yet platforms already arbitrate countless decisions, simply by having and enforcing policies for acceptable behavior.[94] Companies make principled decisions about what is included or omitted in their policies or procedures, and they enact those principles whenever they enforce (or choose not to enforce) them. Instead of clinging to the myth of neutral arbitration, platforms should recognize the power they wield—and the values and principles already evident in the decisions they make every day—and move toward explicitly principled governance.

The philosopher and critical theorist Nancy Fraser proposes that accountability for harms involves "seek[ing] institutional remedies for institutionalized harms."[95] Principled governance requires transparency, accountability, and opportunities for appeal[96]—values which are central in theories of procedural justice,

---

[93] See the field of science and technology studies for a review of how science is produced, e.g., BRUNO LATOUR & STEVE WOOLGAR, LABORATORY LIFE : THE SOCIAL CONSTRUCTION OF SCIENTIFIC FACTS (1979); GEOFFREY C. BOWKER & LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES (1999).
[94] GILLESPIE, *supra* note 2.
[95] Fraser, *supra* note 84.
[96] The Santa Clara Principles, supra note 2.

or the notion that fair and transparent decision-making processes result in more equitable outcomes and, in turn, more cooperative behavior.[97] Some social media companies have begun to respond to public concern about procedural fairness, implementing systems for appealing content removal decisions and experimental initiatives like Facebook's controversial Oversight Board, a group of experts with the authority to overturn a selection of appealed content moderation decisions.[98]

However, principled governance also requires interrogating the limitations of concepts like fairness, despite—or because of—their deep embeddings in many justice systems. Power differences explain why concepts like fairness can overlook injustices: fairness maintains power differentials because it locates the source of problems within individuals or technologies instead of as systemic and contextual inequities.[99] As such, we propose that social media governance must be principled rather than neutral, and that a principled approach requires platforms to reckon with their role in enabling, or magnifying, structural injustices.

From Equality to Equity

Social media companies have traditionally built their policies and procedures around equality, or the notion that all people deserve equal treatment. But equal treatment—which many people may, on its face, consider to be fair—is typically engaged on an individual level, rather than contextualized in a larger system of sociohistorical relationships and systemic injustice. In other words,

---

[97] BRADFORD ET AL., *supra* note 2. *See* Badeie et. al, this issue.
[98] For an in-depth analysis see Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L. J. 2418 (2019).
[99] Anna Lauren Hoffmann, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFO., COMMC'N & SOCIETY 900 (2019).

while equality aims to promote justice and fairness, it can only work if everybody starts with the same resources and needs. In practice, an equality-based approach—when applied to inequitable systems—only serves to uphold existing systems of oppression and perpetuate systemic inequality, such as racism and transphobia. Most (if not all) social media companies apply their policies using policies of equality, thereby perpetuating equalities rather than remediating them.

For example, Facebook's Community Standards define hate speech as "a direct attack"[100]—described as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation—"against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease." The policy is delineated by different types of attacks, but it applies equally to all groups: a dehumanizing statement against men (e.g., "Men are trash") is treated the same as a dehumanizing statement against women (e.g., "Women are trash"), despite structural sexism (i.e., systematic gender inequality, one manifestation of which is the wage gap[101]).

Thus, while "equal treatment" may seem appropriate on an individual level, it obscures—and ultimately perpetuates—existing inequalities at the structural level. Women, queer people, people of color, dissidents, religious minorities, lower caste groups, and other

---

[100] *Facebook Community Standards on Hate Speech*, FACEBOOK, https://www.facebook.com/communitystandards/hate_speech (last visited Apr. 4, 2021).

[101] Nikki Graf, Anna Brown & Eileen Patten, *The Narrowing, but Persistent, Gender Gap in Pay*, PEW RESEARCH CENTER (Mar. 22, 2019), https://www.pewresearch.org/fact-tank/2019/03/22/gender-pay-gap-facts/.

groups are disproportionately affected by online harassment [102], particularly when those identities intersect (e.g., a Black trans woman). Why would we expect social media companies to police harassment of these groups with the same fervor—or to detect it at the same volume—as the less frequent and typically lower-severity harassment of their socially-dominant counterparts? Instead, we argue that social media governance should prioritize equity, or the fair distribution of benefits, resources, or outcomes. This is best understood as a question of distributive justice: whereas equality mandates that everyone is given the same resources or opportunities, an equitable approach recognizes that individual circumstances may require uneven distribution in order to ultimately reach an equal outcome.

Because social differences between people (e.g., race) shape what kinds of harm they might experience (e.g., racism), appropriate responses to harm should be interpreted in the broader cultural and social contexts in which the harm occurred. Although behaviors like online harassment manifest as interpersonal conflict, social media platforms contribute to and perpetuate inequities that result in disproportionate harm to vulnerable populations. To successfully recognize and repair harm, social media companies must first address their role in enabling and exacerbating existing structural injustice.

---

[102] Shawna Chen, Bethany Allen-Ebrahimian, *Harassment of Chinese dissidents was warning signal on disinformation*, AXIOS (Jan. 12, 2021), https://www.axios.com/chinese-dissidents-disinformation-protests-7dbc28d7-68d0-4a09-ac4c-f6a11a504f7c.html; Maeve Duggan, *1 in 4 black Americans have faced online harassment because of their race, ethnicity*, PEW RESEARCH CENTER (Jul. 25, 2017), https://www.pewresearch.org/fact-tank/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/; Duggan, *supra* note 35; Emily Vogels, *The State of Online Harassment*, PEW RESEARCH CENTER (Jan. 13, 2021), https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

From Content to Behavior

Social media companies currently evaluate potentially harmful behavior purely at the content level—that is, content moderators are asked to consider the specific words used in a given post or comment, divorced from contextual factors such as who the author is; who the audience or target is; what the relationship between the author and their audience is, and so on.

While human content moderators will intuit some amount of context from the content itself—for example, a tweet that contains profanity but also a playful emoji may be interpreted as banter between friends—algorithmic (i.e., computational) moderation still cannot. Scaled moderation relies almost exclusively on natural language processing and other machine learning techniques; a typical supervised learning model will be trained on a broad corpus of content and produce blunt, binary judgments—e.g., violating or not violating; hate speech or not hate speech—based on how closely an object resembles the training data set. This results in enforcement outcomes which are almost entirely based on isolated pieces of content, devoid of the sociohistorical context in which they were produced.

Complex and inherently social behaviors like online harassment cannot be understood separate from the context in which they occurred. While the core experience of online harassment may be largely universal across regions and cultures, how people experience harm may vary by individual, context, and culture. For example, non-consensual sharing of intimate images is an intense invasion of privacy regardless of the target's location—but for women in Pakistan or Saudi Arabia, an intimate image could bring shame to an entire family, creating additional consequences and intensifying an already acutely harmful experience.

A focus on behavior allows for more nuance in what sanctions are applied to potential violators. While dominant models of social media governance typically favor blunt punishments that escalate in severity (e.g., limiting a violator's account privileges for one day after their first violation, three days after a second infraction, and so on), this approach has several limitations. First, applying the same punishment to all policy violators, regardless of the infraction, collapses a wide range of behaviors into a binary determination of "violation" or "no violation." In addition to creating uncomfortably disproportionate outcomes—someone who reacts with justifiable hostility to an instance of racism, for example, will endure the same punishment as someone who posts something racist—this approach does not allow for accountability that more appropriately addresses the root cause of specific behaviors.

Content-centric approaches to social media governance also do not account for differences in what motivates individuals to participate in abusive behavior. While the resulting harm is ultimately the same regardless of the perpetrator's intent, considering the underlying motivation for a behavior allows for more strategic and targeted interventions that may reduce the likelihood of reoffense. For example, a user who is new to a specific social media site may benefit from educational interventions that help the user acclimate to platform rules and norms; a user who engages in retributive harassment is likely aware that they are violating a rule. That user could be prompted to report the person they are seeking to sanction instead.

Finally, content removal is an inherently reactive governance strategy; by the time a post is reported to or reviewed by the platform, it has likely already caused significant harm. Reactive governance is a losing game: users are producing content much faster than platforms can moderate it, no matter how many

algorithms they build or moderators they hire. Shifting focus from removing individual pieces of content toward understanding and addressing the underlying behaviors will allow social media platforms to become more proactive in their governance, implementing interventions that discourage harmful behaviors before they manifest on the platform.

From Retribution to Rehabilitation

While criminal justice is an accessible metaphor, it is not a desirable approach to social media governance for a variety of reasons—not least because it privileges a carceral approach that focuses on punishing, rather than rehabilitating, offenders. Retributive governance seeks to restore justice by giving the offender their "just deserts," or a punishment proportional to the offense. While this approach accounts for the severity of harm inflicted, it does nothing to redress the harm itself—in other words, it focuses on the perpetrators of harm, with little to no consideration for the experiences of those who were harmed.

In order to appropriately repair harm, we must first transform social media governance from a system of retribution toward one of accountability. We can draw inspiration from principles of restorative justice, which first asks the injured party to identify their desired path forward. Often, this includes asking the offender to take active accountability for the harm they have caused. Rather than incarcerating offenders, a restorative justice approach seeks to rehabilitate offenders and reintegrate them into the community, reducing the likelihood of recidivism.

This is not to say that punishment is never appropriate. A focus on rehabilitation over punishment allows platforms to better distinguish users who intend to cause harm from those who don't—a distinction many community members already make, particularly

in smaller online communities where moderators frequently interact directly with other users.[103] While good intentions may not lessen any resulting harm, they help indicate an appropriate response. On social media, as in offline contexts, a small number of frequent offenders produce a disproportionate amount of violations; some motivated by extrinsic factors (e.g., financial gain) and others by behaviors associated with violence and manipulation.[104] When offending users are given opportunity to correct and make amends for their behavior, more severe penalties, such as IP address-based or device-level bans, can eventually be applied with more legitimacy. This allows platforms to lessen the intensity of negative experiences caused by incorrect enforcement decisions (e.g., model false positives) while also ensuring that extreme offenders are met with swift punitive responses—resulting in safer, more equitable online spaces.

From Authority to Community

We encourage social media platforms to transition away from paternalistic, top-down models of governance in favor of giving communities more control over their own experiences. One reason for this approach is practicality: these are extremely difficult problems that will take years, if not decades, to solve. Online audiences are disparate and often invisible—even to platforms themselves—making it difficult to reliably assess the targets, scope,

---

[103] Blackwell et al., *Harassment in Social Virtual Reality: Challenges for Platform Governance*, *supra* note 80.

[104] Extensive studies by Neumann and colleagues suggest that about 1% of the population exhibits what has been called psychopathy; however, the psychopathy diagnosis has been contested as overlooking a range of experiences (e.g., disabilities that may falsely present as psychopathy) and should be considered cautiously. Craig S. Neumann & Robert D. Hare, *Psychopathic Traits in a Large Community Sample: Links to Violence, Alcohol Use, and Intelligence*, 76 J. CONSULTING & CLINICAL PSYCH. 893 (2008); Craig S. Neumann et al., *Psychopathic Traits in Females and Males Across the Globe*, 30 BEHAV. SCIS. & L. 557 (2012).

and severity of harms. Platforms are often responsible for evaluating interactions without the necessary context; even when context is available, it is incredibly hard, if not impossible, to evaluate consistently at the scale required to train a machine learning model. We also can't rely on human moderation alone; while automated enforcement has significant limitations, content moderation is incredibly taxing on workers, who spend every day reviewing the worst of humanity for extremely low wages.

Beyond the practicality of more bottom-up, community-driven governance, giving communities increased agency ultimately reduces harm, both by empowering people to exert control over their own experiences and by creating opportunities for more nuanced, individualized interventions. Increased user agency also helps mitigate the challenges of platforms' traditional, "one-size-fits-all" approach to global governance: when communities experiencing harm have control over their experiences on the platform, they can decide what justice looks like on their own terms.

Finally, the transition from authority to agency is necessary for decentralizing the incredible amount of power social media companies now wield. Current approaches to social media governance are fundamentally authoritarian; companies exert total control over their content moderation processes, with little to no transparency into how policies are developed, how moderators make decisions, how algorithms are trained, and every other facet of this incredibly complex ecosystem. Social media platforms exist to serve social functions: relationship-building, free expression, collective organizing. We deserve radical transparency into how this small handful of American companies is choosing to govern what are now our primary social spaces.

CONCLUSION

Despite early optimism about social media's democratic promises, social media platforms have enabled abuse and amplified existing systemic injustices. Models of governance that may have sufficed in early, online communities are ineffective at the scale of many contemporary platforms, which largely rely on obscure but powerful automated technologies. Failures to effectively govern platforms manifest in severe consequences for social media users, including psychological distress, physical violence, and the continued suppression of non-dominant voices. Unfortunately, platforms' reproduction of punitive models of governance focus on removing offenders rather than repairing harm. We argue that platforms are obligated to repair these harms, and that doing so requires reimagining governance frameworks that accommodate a wider range of harms and remedies. We propose a set of governing principles to better equip social media companies for accountability to their users.