Transforming Social Media Governance With Applied Theories of Justice

by

Lindsay Blackwell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2025

Doctoral Committee:

Professor Sarita Schoenebeck, Co-Chair
Professor Cliff Lampe, Co-Chair
Professor Nicole Ellison
Associate Professor Alice Marwick, University of North Carolina at Chapel Hill
Associate Professor Sun Young Park

Lindsay Blackwell

lblackw@umich.edu

ORCID iD: 0000-0003-1553-5596

Dedication

To the survivors.

Acknowledgements

This dissertation is made possible by my mentors, colleagues, and friends.

First, and most fervently, thank you to my advisors, Dr. Sarita Schoenebeck and Dr. Cliff Lampe, who saw in me a potential that I hadn't yet seen in myself, and who have relentlessly supported my pursuit of this degree. Sarita: thank you for your generosity and steady guidance. You taught me how to be a researcher, and you've proven that tremendous accomplishment is not in conflict with, but instead made possible by, investment in life's other areas. I hope to carry with me your thoughtfulness and patience. Cliff: thank you for your mentorship, your impartial ear, and your self-awareness. You taught me how to be an academic, and you modeled the importance of serving our community and building relationships with the people it comprises. You are a tireless advocate and a tiresome person, and for both I am forever grateful.

Thank you to my committee members, Dr. Nicole Ellison, Dr. Alice Marwick, and Dr. Sun Young Park. Nicole: your rigor, humor, and leadership are without parallel, and your mentorship and friendship both foundational to my success. Alice: you planted the seed that would eventually blossom into my research agenda. Thank you for giving me something to aspire to. Sun Young: I met you late in my journey, and my work is better for it. Thank you for bringing your expertise to my committee, and for your enthusiasm and encouragement.

Aaron Abbarno, Mark Ackerman, Nathaniel Adam, Katie Adamsky, Yosef Adiputra, Ashwin Agarwal, James Alexander, Andrea Alexandre, Sanna Ali, Mena Alsrogy, Morgan Ames, Tawfiq Ammari, Nazanin Andalibi, Spencer Anthony, Vidhya Aravind, Carolina Are, Ahmer Arif, Zahra Ashktorab, Seyram Avle, Riad Azar, Savannah Badalich, Kevin Baker, Aruna Balakrishnan, Renata Barreto, Joe Bayer, Nancy Baym, Natalie Bazarova, Edward Beasley, Rachel Bell, Luca Belli, Rosie Bellini, Emily Bender, Susan Benesch, Michael Bernstein, Umang Bhojani, Noah Bialos, Nicole Bjelica, Kathryn Blackwell, Jessica Bodford, Jean Hardy Bohaczek, Nicole Bonoff, Richelle Booc, Erin Brady, Robin Brewer, Jed Brubaker, Amy Bruckman, Moira Burke, Alan Byrne, Robyn Caplan, Shane Caraveo, Brandon Carlisle,

Tim Carmody, Bryan Carroll, Ana Villar Casas, Ulrick Casseus, Stevie Chancellor, Eshwar Chandrasekaran, Clarissa Chang, Ronnie Chen, Tianying Chen, Rumman Chowdhury, Jill Christ, Connie Chung, Elizabeth Churchill, Graham Clenaghan, Terri Conley, Alana Conner, Mike Conner, Olivia Conti, Adam Cowden, John Mark Cuadro, Danielle Czarnecki, Ryan Dansby, Julia DeCook, Nicola Dell, Michael Ann Devito, Jill Dimond, Anne Disabato, Autumn Dishnow, Brian Doctor, Lynn Dombrowski, Evelyn Douek, Joe Drennan, TJ Du, Maeve Duggan, Michael Dworsky, Elizabeth Dwoskin, Brianna Dym, Matthias Eck, Paul Edwards, Andrea Effgen, Natasha Elliott-Deflo, Emily Falk, Gina Marie Falk, Umer Farooq, Michael Feldman, Casey Fiesler, Tom Finholt, Colin Fitzpatrick, Nick Fohs, Liz Fong-Jones, Brittany Forks, Andrea Forte, Sarah Fox, Tara Franz, Colin Fraser, Megan French, Milind Ganjoo, Patricia Garcia, Mariel García-Montes, Emma Gardiner, Pierre Garrigues, Drew Gassaway, Stuart Geiger, Anna Gibson, Eric Gilbert, Sarah Gilbert, Tarleton Gillespie, David Gillis, Alana Glassco, Tonei Glavinic, Ruben Gomez, Kate Grandprey-Shores, Kathryn Grant, Kishonna Gray, Datrianna Green-Meeks, Mary Grey, Dylan Griffin, Brady G'Sell, Tamy Guberek, Shion Guha, Rachelle Mae Guinto, Dulshani Gunawardhana, Amy Guth, Noémie Hailu, Oliver Haimson, Ted Han, Jeff Hancock, Mark Handel, Alex Hanna, Christina Harrington, Del Harvey, Caitlin Hayward, Margaret Hedstrom, Libby Hemphill, Daniel Herron, Mar Hicks, Robyn Hightower, Alexis Hiniker, Marlene Hirose, Anna Lauren Hoffmann, Julie Hollek, James Hooks, Shuangyi Hou, Youyang Hou, Jenn Huang, Mary Beth Hunzaker, Jevan Hutson, Krzysztof Ignasiak, Jane Im, Lilly Irani, Mike Isaac, Vaishnavi J, Morry Jackson, Shagun Jhaver, Lifeng Jia, Aaron Jiang, Jasmine Jones, Frantz Joseph, Nicholas Judd, Younghee Jung, Sanjay Kairam, Cindy Lin Kaiying, Bridget Kaluzny, Julia Kamin, Frank Kanayet, Arcadiy Kantor, Sema Karaman, Matthew Katsaros, Harmanpreet Kaur, Jofish Kaye, Brian Keegan, Wendy Kellogg, Kirin Khan, Vera Khovanskaya, Charles Kiene, Erin Kissane, Viktor Kjellström, Shamika Klassen, Kate Klonick, Michael Kolhede, Kolina Koltai, Alekhya Kommasani, Paige Kordas, Jenny Korn, Marion Lang, Liliya Lavitas, Alex Leavitt, Mariah Leavitt, Beth Leber, Allen Lee, Liz Lee, Paul Lee, Christine Lehane, Janni Lehrer-Stein, Amanda Lenhart, Becca Lewis, Terry Lin, Silvia Lindtner, Eden Litt, Kat Lo, Freia Lobo, Bill Locke, Brandon Locke, Ela Locke, Michelle Locke, Micah Loewinger, Vince Lozada, Caitie Lustig, Arlene Macayan, Francine Romine MacBride, Jeffrey MacKie-Mason, Fadzai Madzingira, Brett Major, Amy Maoz, Megh Marathe, Andrea Marchesini, Liz Bright Marquis, Allan Martell, Azza El Masri, J. Nathan

Matias, Adrienne Massanari, Courtney McCluney, Shilpa Mody, Wendy Moncur, Andrew McDiarmid, Lisa McInerney, Adrianna McIntyre, Tracey Meares, Amanda Menking, Danaé Metaxa, Erika Miley, Andrew Miller, Adriana Miramontes, Andrés Monroy-Hernández, Christopher Moon, Eider Moore, Tsubasa Morioka, Josh Morrison, Carol Moser, Kevin Munger, Alice Munyua, Tyler Musgrave, Lisa Nakamura, Daniel Nazer, Gina Neff, David Nemer, Logan Newkirk, Mark Newman, C. Thi Nguyen, Peipei Nie, Conrad Nied, Safiya Noble, Mick O'Flaherty, Carolyn O'Hara, Olivia Opdahl, Venessa Paech, Rita Parada, Eli Pariser, Jessica Pater, Desmond Patton, Katy Pearce, Claire Pershan, Jerrel Peterson, Kathie Pham, Karolina Pietkiewicz, Darya Pilram, Ignacio Contreras Pizarro, Alessio Ponzio, Nat Poor, Miriam Posner, Stephen Potter, Edel Quinn, Lucracia Ray, Olivia Redfield, Elissa Redmiles, Nikkia Reveillac, Isabelle Rice, Tylea Richard, Megan Riley, Ryan Ritter, Sarah Roberts, Danae Ford Robinson, John Robinson, Meghan Rolfe, Daniela Rosner, Ashley Roybal-Reid, Jennifer Rubin, Cass Wilkinson Saldaña, Niloufar Salehi, Christian Sandvig, Christine Sargent, Pradeep Kumar Sasidharan, Colby Sato, Morgan Klaus Scheuerman, Zoë Schiffer, Raz Schwartz, Carol Scott, Joseph Seering, Jessica Seguis, Jessica Webster Sendra, Zayba Shaikh, Natalie Shaw, Juliet Shen, Cammie Shibata, John Simas, Caroline Sinders, Malika Singh, Kyera Singleton, Jan Smole, Matthew Snyder, Sarah Sobieraj, Luca Soldaini, Olivia Solon, Nneka Soyinka, Simon Spicer, Katta Spiel, Jeff Stade, Kathleen Stanko, Kate Starbird, Luke Stark, John Starr, Risa Stein, Peter Stern, Angelika Strohmayer, Kaitlin Sullivan, Sammy Sussman, Nicholas Suzor, Moiz Syed, Catherine Teitelbaum, Jacob Thebault-Spieker, Krishna Thiagarajan, John Charles Thomas, Michaelanne Thomas, Andrea Thomer, Adam Tishok, Sam Toizer, Catalina Toma, Ruth Toner, Fortuno Manuel Torres, Manuela Travaglianti, Penny Triệu, Rebekah Tromble, Janice Tsai, Joe Tullio, Tom Tyler, Kristen Vaccaro, Aparajitha Vadlamannati, Tiffany Veinot, Jacqueline Ryan Vickery, Sarah Vieweg, Jessica Vitak, Kimberly Voll, Andreas Wagner, Brian Waismeyer, Jazz Walker, Jen Weedon, Sarah Myers West, Tarin Whitehouse, Ben Whitelaw, Amy Whitney, Elle Whittaker, Andrew Williamson, Charlotte Willner, Elizabeth Wingrove, Pamela Wisniewski, Shoshana Wodinsky, Yvette Wohn, Christine Wolf, Cynthia Wong, Emily Wood, Mary Woolston, Graham Wright, Diyi Yang, Sarah Ybarra, Jillian York, Andrea Zeller, Amy Zhang, Lotus Zhang, Xuan Zhao, my research participants, and surely others I've neglected: Thank you, for everything. It's been a pleasure to be in your company.

I am better for knowing all of you.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation explores ways in which social media governance can be improved through the practical application of theories and principles of justice. Three empirical research studies demonstrate that current platform governance practices exacerbate the very harms they seek to mitigate, due in large part to the influence of punitive criminal justice models that are structurally incapable of producing equitable governance outcomes. Informed by transformative justice, social media governance can be reframed from a method for control to a practice of collective care, requiring a fundamental restructuring of the systems and logics that currently govern online spaces.

Chapter 1 Introduction

Despite a rich history of scholarship on the regulation of behavior in online spaces (e.g., Dibbell, 1993; Donath, 1999; Lampe & Resnick, 2004; Bruckman et al., 2006), attempts at governing contemporary social media platforms—such as Facebook, Twitter, and Youtube—have largely failed. Social media companies have typically attempted to govern their platforms by developing and enforcing individual policies for use, including what behaviors are and are not acceptable; however, enforcing these policies at a global scale has proven to be an impossible challenge (Gillespie, 2018; Roberts, 2019). Many social media users are not aware that rules exist; a user is often first exposed to platform rules when they have violated one, often inadvertently (Tyler et al., 2025). Other users may have an ambient awareness of the existence of site policies, but few can articulate specific rules or examples of inappropriate behavior—a problem intensified by the lack of transparency social media companies provide into their specific enforcement practices (Suzor, 2019).

The success or failure of social media governance is dependent on three primary factors: user behavior, platform affordances, and company procedures. How an individual user chooses to behave is primarily influenced by *social norms*, or the values, customs, stereotypes, and conventions they encounter through their interactions with others (Sherif, 1936). Social norms differ from codified laws in that they are socially negotiated and learned through social interactions; by understanding how most people behave in a given situation, a social actor can more quickly decide how to behave themselves (Cialdini et al., 1990). Both formal sanctions (including formal policies and laws) and informal sanctions (such as shame, ridicule, disapproval, or ostracism) facilitate the regulation of non-normative—or deviant—behavior, which in turn establishes expectations for appropriate conduct (Erikson, 1966). Deviance is socially constructed, and perceptions of what is or is not considered deviant behavior shift over time in accordance with dominant cultural norms (Becker, 1963).

Though scholarship across disciplines has long been concerned with the influence of social norms—and in particular, with the sanctioning of deviant behaviors—emerging technology contexts significantly alter the ways in which social norms are created, violated, and enforced. For example, social media sites, which afford scalability, speed, and persistence, enable groups of users to coordinate attacks on individual targets—one of several behaviors frequently described as *online harassment*. Over 40% of American adults have personally experienced online harassment, including being called offensive names, being purposefully embarrassed or physically threatened, or being stalked or harassed for a sustained period of time, with the majority of targets (75%) reporting experiences on social media specifically (Vogels, 2021). More severe online harassment experiences—including physical threats, stalking, and sexual harassment—are increasingly common, with 25% of American adults reporting these experiences in a 2020 survey, compared to just 18% in 2017 (Vogels, 2021).

Theories of social deviance are often intertwined with theories about crime, with individuals who willingly violate formal rules and laws (e.g., criminals) considered to be "deviants." Given the historical relationship between deviance and criminality, it is perhaps unsurprising that most large social media companies have adopted policies and procedures for governing undesirable behaviors that largely mimic existing systems of criminal justice, wherein people who violate formal rules or laws are held accountable for their actions via a range of potential interventions. In Western countries, this is typically achieved through the application of penalties and punishments. Indeed, most American social media companies have developed formal, written rules—for example, Meta's "Community Standards"[1]—describing specific unacceptable behaviors and associated penalties. Users who are determined to have violated these rules are subject to a variety of sanctions, including the permanent removal of a user's content or account.

This dissertation explores ways in which social media governance can be improved through the practical application of justice theories and principles. First, I review relevant literature on the creation and violation of social norms, including how norm violations (i.e., social deviance) are both implicitly and explicitly sanctioned. I then provide an overview of the history of behavioral regulation in online communities—including normative, distributed,

[1] http://transparency.meta.com/policies/community-standards/

algorithmic, and retributive paradigms (Schoenebeck & Blackwell, 2021)—and various interpersonal and societal harms resulting from ineffective platform governance. The literature review concludes with a summary of five major theories of justice (procedural, retributive, restorative, distributive, and transformative justice) and raises several questions about the successful application of justice theory in online spaces. I return to these questions in the discussion, leveraging the learnings from three empirical studies to discuss how we might achieve more equitable social media governance.

The first study (Blackwell et al., 2017) investigates problems of classification inherent in the detection, evaluation, and sanctioning of online harassment. While natural language processing and other machine learning techniques are promising approaches for identifying abusive language at scale, they fail to address structural power imbalances perpetuated by automated labeling and classification. Similarly, platform policies and reporting tools are designed for a seemingly homogenous userbase, and as such they do not adequately address individual experiences and systems of social oppression. This research examines systems of classification enacted by technical systems, platform policies, and users to demonstrate why explicitly labeling content as harmful is critical for surfacing community norms around appropriate user behavior—and why fully addressing online harassment requires the ongoing integration of vulnerable users' needs into platform design and governance.

The second study (Blackwell et al., 2018) applies a specific theory of criminal justice—retributive justice, or the application of penalties that are proportionate to the specific offense—to the perception of online harassment as justified or deserved. A retributive justice framework can help us better understand why online harassment occurs, as well as what motivates the decisions of moderators and bystanders, who may only choose to act against (e.g., by flagging or reporting) users whose actions they do not consider to be justified. This research demonstrates that punitive models of criminal justice break down in online environments, where offenders can hide behind anonymity and lagging legal systems, and where users may instead turn to their own moral codes to sanction perceived offenses. This study helps us to better understand the challenges of implementing justice at scale, a point I return to in the discussion.

The third and final study (Blackwell, 2025) broadens current understandings of social media governance by examining the lived experiences of practitioners who enact, examine, and engage with scaled content moderation systems across various professional contexts. Through a series of participatory design workshops with content moderation professionals—from part-time content moderators and university researchers to corporate vice presidents—I produce a more intimate understanding of the complex landscape of people, practices, and politics that ultimately determine how contemporary social media platforms are governed. By examining the current state of scaled content moderation through a worker-centric lens, this study seeks to identify the underlying logics that construct and sustain social platform governance and articulate aspirational, worker-centered visions for more equitable technology futures.

I conclude by returning to transformative justice, which aims to address and prevent harm by transforming the specific social conditions that create and perpetuate injustice. While frameworks of procedural, retributive, restorative, and distributive justice offer useful lenses for understanding and evaluating current platform governance practices, transformative justice provides a more radical foundation for reimagining what social media governance could become. Applying a transformative justice lens to social media governance requires a fundamental reimagining of how we define and respond to online harm, including challenging the structural incentives that motivate technological scale. Truly transforming online governance will require attending to global power asymmetries, to produce technologies that are appropriately responsive to the diverse needs, values, and identities of the communities they serve.

Chapter 2 Literature Review

This literature review summarizes key theoretical frameworks and empirical studies concerning the governance of behavior in online social spaces and relevant theories of justice, as a foundation for better contextualizing the three empirical studies that follow.

2.1 Governing human behavior

The governance of human behavior is a foundational concern across the social sciences. Human behavior is regulated at individual, social, and institutional levels, through a variety of formal and informal mechanisms for defining—and enforcing—the boundaries of acceptable conduct. At its core, behavioral regulation involves the shaping, constraining, or enabling of individual actions within a given social context.

While law and policy are often the most visible and explicit instruments of regulation, an individual's actions are shaped by their broader social context. Lessig (2006) proposes that human behavior is regulated by four major forces: law (or formal rules, often enforced at the institutional level), norms (or informal rules, shaped by our social experiences), markets (economic incentives and disincentives), and architecture (design constraints, including the design of technical systems). These mechanisms do not operate in isolation, but instead dynamically produce acceptable patterns of social conduct. In online environments, Lessig (2006) argues that "code is law"—the technologies we use are designed in certain ways, and embed particular values, that ultimately govern our behavior. While traditional laws require enforcement (e.g., by law enforcement officers, courts of law, and so on), code enforces itself—instantly, invisibly, and often without recourse (Lessig, 2006). As a result, technology companies wield considerable power over public discourse, with limited transparency or accountability.

2.1.1 Social norms

Social norms, or the unwritten codes of conduct that influence behavior at both the individual and societal level (Chung & Rimal, 2016), are the foundation of individual behavior and social interactions, both in the physical world and in online spaces. Social norms differ from codified laws in that they are socially negotiated and learned through social interactions. Social norms—such as values, customs, stereotypes, and conventions—are “social frames of reference” that individuals first encounter through their interactions with others, and which later become internalized (Sherif, 1936). Although the effects of social norms are widely studied, less is known about how and why norms emerge. One widely accepted theory posits that norms emerge “to satisfy demands to mitigate negative externalities or to promote positive ones” (Hechter & Opp, 2001). Thus, norms are most likely to emerge when they favorably impact a given community’s goals (Opp, 2001). Suler (2004) refers to the notion of cultural relativity: given the immense variety of online communities, “what is considered asocial behavior in one group may be very à propos in another.”

The concept of social norms is rooted in the idea that individuals’ behaviors are shaped, in part, by the behaviors of others (Chung & Rimal, 2016). Traditional theories of behavior considered learning to be an individual process, governed primarily by reinforcement and punishment; Bandura’s (1971) social learning theory, however, proposed that learning can occur purely through observation or direct instruction. Garfinkel (1967) and his followers advocated for a focus on individual social agency, using the observation of everyday social interactions—what he termed ethnomethodology, influenced by symbolic interactionism—to generate a “common sense knowledge” of our social world. Garfinkel’s work (1967) relied heavily on breaching experiments, or experiments that examine people’s reactions to violations of commonly accepted social norms, to produce accounts of how social actors reason and construct understanding when confronted by an action they cannot immediately explain.

2.1.2 Social deviance

The perceived violation of a social norm is referred to as social deviance. Some scholars regard all deviations from the norm as deviance (Durkheim, 1884); others assert that deviance is only that nonconformity which is not otherwise valued within a given community (Becker, 1963; Erikson, 1966). Theories of social deviance are often intertwined with theories about crime, and people who willingly violate formal rules and laws (e.g., criminals) typically considered to be “deviants.” Classical theories of social deviance and crime were largely naturalistic, asserting that individual or biological factors were largely responsible for deviant behavior (Lombroso, 1911/1876)—and consequently, that the most effective way to discourage potential offenders is through the fear of strict criminal sanctions. Classical theorists felt that the benefits of such an approach—maintaining order and social security—outweighed any potential costs (e.g., the punishment of innocent persons).

Others have theorized deviance as community-dependent and socially constructed, wherein all behaviors have the potential to be defined as deviant. Perceptions of what is and is not considered deviant behavior shift over time, shaped by dominant cultural norms and majority forces (Becker, 1963). For example, beginning in the nineteenth century, homosexuality was defined by medical literature as sexually deviant behavior, largely due to professional medicine's "ideological and material links with upper middle class male society and its consequent role in defining sexuality" (Weeks, 1999). Changing scientific, legal, and ideological understandings of homosexuality and of human sexuality more broadly have substantially—though not entirely—shifted societal perception. Many behaviors throughout human history have at one time been considered socially deviant, including behaviors broadly considered to be prosocial (e.g., political dissent). The importance of social labeling in the emergence and persistence of social norms (Becker, 1963; Weeks, 1999) is critical to understanding societal perceptions of "deviant" behaviors, including the ways in which this label may be leveraged for the oppression of non-dominant groups.

2.2 Contemporary social media governance

Erikson (1966) argues that communities use deviance to establish these boundaries—or rather, that those who misbehave in turn establish community norms, including how formal rules are made, enforced, and broken. As communities develop norms for appropriateness, they enforce those norms through both formal (e.g., policies and laws) and informal (e.g., shame, ridicule, disapproval, or ostracism) sanctions for deviant behavior. Normative appeals are of particular importance in online communities, where group identity is made more salient by the varying degrees of anonymity an online platform affords (Lea & Spears, 1991). As a result, anonymous individuals may be more susceptible to normative pressure (Postmes et al., 2001; Munger, 2017).

Early online communities largely relied on normative appeals for regulating deviance, sometimes selecting community members to enforce certain norms as administrators or moderators (Rheingold, 1993; Schoenebeck & Blackwell, 2021). Schoenebeck and Blackwell

(2021) identify four major paradigms of social media governance, reflecting distinct approaches to moderating online behavior: normative, distributed, algorithmic, and retributive regulation. Normative regulation, the earliest paradigm of online governance (though still applicable today), relies on implicit cultural standards and community norms to guide acceptable behavior, with or without formal enforcement (Pater et al., 2016; Matias, 2019). Although normative regulation enables communities to self-govern according to their own values and priorities, it is most effective in communities with well-defined boundaries, such as individual subreddits (Schoenebeck & Blackwell, 2021). As such, many contemporary social media platforms—such as Twitter and TikTok—are restricted in their ability to successfully regulate behavior through normative means.

Distributed regulation—popularized by platforms such as Slashdot in the early 2000s (Lampe & Resnick, 2004), though still used by contemporary platforms, such as Reddit—shifts responsibility to users, delegating the responsibility of governance across loosely connected actors. Under this model of governance, communities leverage scaled feedback mechanisms (such as upvotes and downvotes) to determine collective user preferences. The dominant governance paradigm for modern social media platforms is *algorithmic regulation*, or the use of automated systems to detect and remove content determined to violate particular rules, often with limited transparency or recourse (Schoenebeck & Blackwell, 2021). Because contemporary social media platforms operate at global scale, machine learning techniques such as natural language processing allow for systematic evaluation of large quantities of data—though accurate machine detection of nuanced human behavior is "challenging at best" (Schoenebeck & Blackwell, 2021). Instead, harmful content frequently evades detection, and permissible content may be inadvertently removed.

2.2.1 Scaled content moderation

The promise of scaled content moderation has been increasingly undermined as contemporary social media platforms repeatedly fail to adequately prevent harm, both to individuals and society at large. When social media companies fail to appropriately govern their platforms, a fourth paradigm of governance emerges: *retributive regulation*, wherein "social media users aspire to enforce justice themselves," often resulting in extreme or disproportionate impacts for perceived offenders (Blackwell et al., 2018; Schoenebeck & Blackwell, 2021).

Interdisciplinary scholarship has highlighted numerous ways in which scaled moderation systems fail, both due to insufficient technical systems as well as the structural, economic, and epistemological assumptions embedded within them (Gillespie, 2018; Roberts, 2019; Gorwa et al., 2020). Platform design decisions profoundly shape what kinds of governance are possible and how values are operationalized at scale. Platforms must consider both preventative and reactive strategies for mitigating misbehavior in their communities (Bruckman et al., 2006), and still, some users may actively violate site policies in their efforts to intentionally damage a community or its members (Shachaf & Hara, 2010). Individual users who wish to regulate their own exposure to misbehavior may make use of blocklists and other filters, which reduce the visibility of known offenders but do not restrict their future conduct (Donath, 1999). Moderation tools and other governance processes encode specific assumptions about risk, harm, and legitimacy (Gray & Suri, 2019; Seering, 2020). These systems frequently prioritize efficiency and throughput over nuance and care, flattening complex social dynamics into binary content removal decisions.

Efforts to prevent antagonistic online behaviors through the creation and enforcement of legal policy have also largely failed. Despite increasing public pressure to discourage and even limit speech that is perceived to be discriminatory, threatening, or otherwise counterproductive to democratic goals (Grimmelmann, 2014), many social media platforms operate under the guise of neutrality (Gillespie, 2010), in part because of the limited liability enjoyed by information providers under current regulatory systems (e.g., under Section 230 of the Communications Decency Act). Online communities, social network sites, and news sites adopt individual policies for use, including what behaviors are and are not acceptable; however, enforcing these policies can be problematic, if not impossible, at scale (Blackwell et al., 2017; Gillespie, 2018). Legal mechanisms have also failed to adequately support those targeted by harmful behaviors online: online harassment targets have found little (if any) recourse by legal means, as authorities often do not consider online threats to be serious, or lack the resources to appropriately pursue them (Merlan, 2015).

Attempts to scale traditional online moderation practices—e.g., appointing a small set of users with special privileges that enable them to remove undesirable content—have resulted in a complex ecosystem of platform governance that relies heavily on what Sarah Roberts (2016) calls *commercial content moderation*. Whereas potentially undesirable content might be reviewed by a volunteer moderator in a smaller online community, commercial content moderation relies on thousands of low-wage workers—typically contracted from companies that now specialize in these practices—to review content reported by users or detected by machine learning algorithms. Workers are exposed to a glut of gut-wrenching content, from vicious racial attacks to gruesome violence, and held to unrealistic performance standards; a typical commercial content moderator will review hundreds of individual pieces of contents or potentially violative accounts on any given day, memorizing extremely complex and constantly evolving policies in order to make an enforcement decision in a matter of seconds.

While this is obviously not a tenable solution, particularly as social media platforms only continue to grow in size, the alternative—automatically detecting and enforcing on content through the use of machine learning algorithms—is equally riddled with problems (Blackwell et al., 2017; Gorwa et al., 2020). Without proactively accounting for social equity in the development of these models, starting with the classification systems a second set of often-contracted workers (Gray & Suri, 2019) are tasked with applying to content to accumulate the volume of data required for training a supervised learning model, systemic biases are reinforced. This results in models that disproportionately detect and enforce against particular types of content, with users from marginalized identity groups enforced against more frequently (Haimson et al., 2021).

2.3 Applied theories of justice

Social media governance reflects broader dynamics of power, identity, inclusion, and control, and systems of behavioral regulation often reflect dominant ideologies, reproducing patterns of social oppression and further marginalizing non-dominant groups. Given the heavy influence of Western conceptions of criminal justice on contemporary platform governance, recent scholarship has explored opportunities for producing more just and equitable online governance through the application of justice principles and practices used in offline contexts (Blackwell et al., 2017; Blackwell et al., 2018; Hasinoff et al., 2020; Schoenebeck et al., 2021; Im et al., 2022; Katsaros et al., 2022; Xiao et al., 2023; Katsaros et al., 2024).

2.3.1 Procedural justice

Procedural justice concerns individuals' perceptions of the fairness of decision-making processes, which is informed not only by the perceived fairness of governance outcomes, but also by the perceived quality of experiences interacting with governance systems—including evaluations of the procedures used to create and implement rules (Tyler, 2006). Perceptions of procedural justice are influenced by four primary factors: voice, neutrality, respect, and trust (Tyler, 2006). Procedural justice is critically important for establishing law enforcement organizations as legitimate authorities, which in turn increases community trust and willingness to comply (Tyler, 2007; Katsaros et al., 2022).

Recent research has explored the impacts of procedural justice in the context of social media governance. Much like in offline contexts, increased transparency in content moderation experiences can increase future rule compliance (Jhaver et al., 2019c; Tyler et al., 2021; Katsaros et al., 2022). For example, users who were provided explanations about why their content was removed were less likely to have future posts removed (Jhaver et al., 2019c). Even less contextual interventions, such as offering users general education about platform rules following content removal experiences, have been shown to reduce future violations (Tyler et al, 2021). General transparency about the specific rules and behavioral expectations of a certain online space is equally important: users who read community rules are more likely to perceive content removals as fair (Jhaver et al., 2019a), and simply posting visible rules increases compliance among new users (Matias, 2019).

2.3.2 Retributive justice

Retributive justice refers to a theory of punishment in which individuals who knowingly commit an act deemed to be morally wrong receive a proportional punishment for their misdeeds, sometimes referred to as “an eye for an eye” (Carlsmith & Darley, 2008; Walen, 2015). Unlike other theories of justice concerned with preventing future wrongdoing or mending conflicts between individuals and communities to the satisfaction of all parties (Wenzel et al., 2007), retributivism is primarily preoccupied with delivering a "just desert" for a morally wrong act (Kant, 1911/1781). For example, in a retributive framework, the death penalty is considered a proportional punishment for an offender who commits murder. Retributive approaches to criminal justice are built around the notion that fear of incarceration or other punishment will deter potential crime.

Retributive justice relies upon the assumption that everyday citizens possess intuitive judgments of “deservingness” that accurately and consistently express the degree of moral wrongdoing of others’ acts. Violations of social norms are sanctioned in a variety of ways, including through formal sanctions (e.g., policies and laws that enforce specific punishments) and informal sanctions such as shame, ridicule, disapproval, or ostracism. But when sanctioning occurs online at scale—aided by affordances such as scalability, speed, and persistence—harassment and other similarly abusive behaviors can become a controversial mechanism for enforcing perceived norm violations, particularly when more formal sanctions are absent or perceived as inadequate (Blackwell et al., 2017; Schoenebeck & Blackwell, 2021). A retributive justice framework helps us better understand and intervene in cases of online harassment perceived to be justified or even deserved.

2.3.3 Restorative justice

An alternative to Western, retributivism-based criminal justice is restorative justice, a philosophical and practical framework for restoring justice through a shared understanding between victim, offender, and community (Wenzel et al., 2007). Whereas retribution is often associated with violence, restoration is rooted in nonviolent approaches to conflict resolution, with the ultimate goal of both protecting communities and encouraging offenders to change their behaviors (Lawrence, 1991). Restorative justice provides a voice to both victim and offender; the victim is encouraged to express a willingness to forgive, and the offender is encouraged to accept responsibility for their actions. Once the victim, offender, and community members reach a shared understanding of the harm caused by the offense and the community values it violated, the victim and offender together determine a suitable punishment (Wenzel et al., 2007).

Given what we know about the antithetical relationship between retributive and restorative theory and the promise of restorative programming in practical settings, exposure to restorative values online could be a promising approach for reducing harmful behavior. In the context of community moderation, a restorative justice framework could allow “space for contestation” (Crawford & Gillespie, 2016); one could envision the design of a community space where members could either defend or condemn certain user behaviors. A limited number of online communities already embody restorative principles in their moderation tactics: for example, League of Legends’ tribunal system, which was active until early 2016, both relieved moderator burden and improved user behavior.

2.3.4 Distributive justice

Technology development has a fraught relationship with social progress. Technology can empower members of non-dominant groups, leveling the proverbial "playing field" and enabling marginalized voices to circulate information widely. However, technology can also amplify existing dynamics of social oppression, particularly when developers do not meaningfully engage with their own subjective biases when building or improving technologies. Noble (2018) details countless examples of racial biases embedded into everyday technological systems: 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." Gender and sexuality discrimination is also prevalent in technology design, from default avatars registering as male silhouettes (Bailey & LaFrance, 2017) to Facebook's ongoing challenges surrounding its "real name" policy and the deactivation of accounts belonging to trans users, drag queens, Native Americans, abuse survivors and others whose identities or account names may be inconsistent with their legal names (Haimson & Hoffman, 2016).

Technologies, never neutral, are instead inextricable from the values and judgments of the humans who build them, however involuntarily those values may be imparted. The reproduction of social inequities in technological contexts can be understood as a question of distributive justice, or the equitable distribution of benefits, resources, or outcomes (Schoenebeck & Blackwell, 2021). Whereas equality mandates that all individuals are given equivalent opportunities or resources, an equitable approach recognizes the impact of systemic oppression on individual circumstances, which may require uneven distribution in order to ultimately achieve equal outcomes. Distributive justice concerns are particularly salient in Silicon Valley, which remains largely driven by competition, innovation, and capital rather than self-reflection and deliberation, a product of what Barbrook and Cameron (1996) describe as a "profound faith" in the democratizing power of technology coupled with "willful blindness" toward social realities such as racism and poverty. Even Howard Rheingold, who once famously described technology as "the great equalizer" (Rheingold, 1991), more recently admitted that Silicon Valley's particular utopianism represents "an idealism inside the framework of capitalism" (Duff, 2016).

2.3.5 Transformative justice

As the empirical research studies in this dissertation will demonstrate, contemporary approaches to social media governance—which rely on punitive interventions and top-down authority—fail to address the structural and relational conditions that enable online harm, including broader social, political, and economic inequality. One potential source of inspiration for achieving more equitable governance is *transformative justice*, which aims to address and prevent violence by transforming the specific social conditions that create and perpetuate injustice (Mingus, 2019; Kaba, 2021). While frameworks of procedural, retributive, restorative, and distributive justice offer useful lenses for understanding and evaluating current platform governance practices, transformative justice provides a more radical foundation for reimagining what social media governance could become.

Transformative justice emerged as a response to the failures of state systems to provide safety or justice for marginalized communities, instead perpetuating violence through carceral logics of surveillance and punishment (Mingus, 2019; Kaba, 2021). Rooted in abolitionist, Indigenous, and Black feminist traditions, transformative justice seeks not to punish wrongdoing, but to address its underlying causes, with the ultimate goal of creating conditions under which harm is less likely to occur. By examining the structural conditions that create and perpetuate violence—including racism, transphobia, poverty, and other intersecting systems of oppression (Crenshaw, 1991)—transformative justice invites us to imagine alternative, community-driven futures, reducing reliance on carceral systems by encouraging communities to take collective responsibility for addressing and preventing harm (Mingus, 2019; Kaba, 2021). Informed by the results of my research, I return to transformative justice in my concluding chapter, where I argue that enacting truly just governance requires sustained commitment to structural change.

Chapter 3 Classification and Its Consequences for Online Harassment: Design Insights from HeartMob[1]

Online harassment is a pervasive and pernicious problem. Techniques like natural language processing and machine learning are promising approaches for identifying abusive language, but they fail to address structural power imbalances perpetuated by automated labeling and classification. Similarly, platform policies and reporting tools are designed for a seemingly homogenous userbase and do not account for individual experiences and systems of social oppression.

This paper describes the design and evaluation of HeartMob, a platform built by and for people who are disproportionately affected by the most severe forms of online harassment. We conducted interviews with 18 HeartMob users, both targets and supporters, about their harassment experiences and their use of the site. We examine systems of classification enacted by technical systems, platform policies, and users to demonstrate how 1) labeling serves to validate (or invalidate) harassment experiences; 2) labeling motivates bystanders to provide support; and 3) labeling content as harassment is critical for surfacing community norms around appropriate user behavior.

We discuss these results through the lens of Bowker and Star's classification theories and describe implications for labeling and classifying online abuse. Finally, informed by intersectional feminist theory, we argue that fully addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms.

[1] Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW).

3.1 Introduction

Roughly four in ten American internet users have personally experienced harassment online, including name-calling, embarrassment, physical threats, stalking, sexual harassment, and sustained harassment (Duggan et al., 2014; Lenhart et al. 2016; Duggan & Smith, 2017). Online harassment can be particularly devastating for marginalized populations, including women of color and lesbian, gay, bisexual, and transgender (LGBT) people (Duggan et al., 2014; Lenhart et al., 2016), limiting their ability to participate safely and equitably in online spaces.

We conducted semi-structured interviews with 18 users of HeartMob, an online platform built by and for people who are disproportionately affected by the most severe forms of online harassment. HeartMob was developed as a grassroots platform under its parent organization, Hollaback!, a non-profit organization dedicated to ending harassment in public spaces. A cornerstone of HeartMob is its community-based approach, which seeks to combat harassment through bystander support—with a focus on amplifying the voices of marginalized internet users. In this work, we build on extensive prior research that has explored misbehavior in online communities (Dibbell, 1993; Donath, 1999; Sproull & Kiesler, 1991) and mechanisms for moderating and mitigating it (Bruckman et al., 2006; Kiesler et al., 2012; Lampe & Resnick, 2004). Our results contribute three insights: first, for harassment targets, labeling experiences as 'online harassment' provides powerful validation of their experiences. Second, for bystanders, labeling abusive behaviors as 'online harassment' enables bystanders to grasp the scope of this problem. Third, for online spaces, visibly labeling harassment as unacceptable is critical for surfacing norms and expectations around appropriate user behavior.

We use Bowker and Star's (2000) classification theories and Becker's (1963) labeling theory of deviant behavior to better explicate the role of power and social oppression in the classification—whether by technical systems, platform policies, or users—of harassing behaviors as normative or non-normative. We discuss the implications of our results for technical approaches to labeling online abuse, which often fail to address structural power imbalances perpetuated by classification.

Finally, we turn to intersectional feminist theory (Collins, 1990; Hooks, 2000; Ahmed, 2017) to further elucidate the limitations of traditional classification systems and to inform potential alternatives. When a classification system is created with dominant values and morals in mind, the needs of marginalized users are neglected. This framework allows us to bring a

CSCW argument to bear on the limitations of current efforts to prevent, manage, or detect online harassment, including platform policies, reporting tools, and machine learning-based approaches. We conclude by advocating for more democratic and user-driven processes in the generation of values that underpin technology systems. Ultimately, centering those who are most vulnerable results in technologies that better address the needs of all users.

3.2 Related work

3.2.1 Online harassment

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). These tactics are often 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 (Ashktorab & Vitak, 2016; Smith et al., 2008) and non-consensual intimate image sharing (also known as "revenge porn"), a form of doxing in which sexually explicit images or videos are distributed without their subject's consent, often by a former romantic partner (Citron & Franks, 2014).

Despite a rich history of research exploring online misbehavior (Dibbell, 1993; Donath, 1999; Sproull & Kiesler, 1991) and community moderation (Bruckman et al., 2006; Lampe & Resnick, 2004), harassment and other forms of abuse remain a persistent problem online. A Pew survey conducted in 2017 revealed that 66% of adult internet users have seen someone be harassed online, and 41% of users have personally experienced online harassment (Duggan & Smith, 2017). Although these behaviors are instantiated using technology (such as social media sites, text messages, or emails), targets of online harassment frequently report disruptions to their offline lives, including emotional and physical distress, changes to technology use, and increased safety and privacy concerns (Duggan & Smith, 2017). Online harassment is also disruptive to everyday life and requires physical and emotional labor from targets, who must spend time reporting abuse in order for platforms to potentially intervene. Indeed, some targets report distractions from personal responsibilities, work obligations, or sleep (Griffiths, 2002; Pittaro, 2007). In addition, online harassment has a chilling effect on future disclosures: Lenhart et al. (2016) found that 27% of American internet users self-censor what they post online due to fear of harassment. A 2017 poll of industry and academic experts revealed fears that harassment and other uncivil online behavior will only get worse, resulting in what many fear could be devastating consequences for free speech and privacy (Rainie et al, 2017).

Although men and women both experience harassment online, women "experience a wider variety of online abuse" (Lenhart et al., 2016) and are disproportionately affected by more serious violations, including being stalked, sexually harassed, or physically threatened (Duggan et al., 2014). Young women are particularly vulnerable to more severe forms of harassment: among all young women surveyed by Pew in 2014 (Duggan et al., 2014), 25% had been sexually harassed online (compared with 6% of all internet users), and 26% had been physically threatened (compared with 8% of all users). People of color are also more susceptible to online abuse: 59% of Black internet users have experienced online harassment. 25% of Blacks and 10% of Hispanics have been targeted with harassment online because of their race, compared with just 3% of white respondents (Duggan & Smith, 2017). Lesbian, gay, and bisexual (LGB) persons also disproportionately experience more serious forms of online abuse: 38% of LGB individuals have experienced intimate partner digital abuse, compared with 10% of heterosexual individuals (Lenhart et al., 2016). LGB persons are more likely "to feel scared or worried as a result of harassment, experience personal or professional harms, or take protective measures to avoid future abuse" (Lenhart et al., 2016).

3.2.2 Technical approaches for addressing online harassment

Recent efforts for addressing online harassment have largely revolved around natural language processing and machine learning techniques for the classification of abusive language, enabling automatic detection and prevention. In one of the earliest published machine learning approaches to harassment detection, Yin et al. (2009) focus on detecting "intentional annoyance" in discussion-style (e.g., MySpace and Slashdot) and chat-style (e.g., Kongregate) communities. Yin et al. (2009) are able to improve a basic supervised model for identifying harassing posts through the addition of contextual features (the context in which a given post occurs) and semantic features (for example, foul language combined with the use of second-person pronouns). More recently, Chandrasekharan et al. (2017) draw on large-scale, preexisting data from other online communities (4chan, Reddit, Voat, and MetaFilter) to generate a computational model that, when applied to an unrelated community, can identify abusive behavior with 75% accuracy (and 92% accuracy after the model is trained on 100,000 human-moderated posts from the target community). Similarly, Wulczyn et al. (2017) use a supervised classifier trained on 100,000 human-labeled Wikipedia comments to automatically identify over 63 million personal attacks across the platform. Other automated attempts to detect abusive language online include Google and Jigsaw's Perspective project, an API that uses machine learning models to assign a "toxicity score" to a string of input text (Hosseini et al., 2017).

3.2.2.1 Limitations of automated approaches

Though harassment detection approaches have improved dramatically, fundamental limitations remain. As Hosseini et al. (2017) demonstrate with Google's Perspective, automatic detection efforts are easily outmaneuvered by "subtly modifying" an otherwise highly toxic phrase in such a way that an automated system will assign it "a significantly lower toxicity score." Yin et al. (2009) also note that too many spelling errors render the sentiment features of their detection model ineffective. Automated approaches to sanctioning, such as Twitter's 'time out,' frustrate users who perceive them as opaque or unfairly applied. In one case, a transgender woman was sanctioned for including the phrase "Fuck you" in a tweet directed to @VP, the government account of Vice President Pence (Warren, 2017). This event, which some users perceived as a limitation of a citizen's right to criticize government administrations, drew outrage from users who had unsuccessfully reported racist content from white supremacy groups using the same platform—demonstrating the difficulty of imparting social nuance into the classification of harassing behaviors at scale.

3.2.3 Classification has consequences

Fundamentally, problems that arise from technical approaches to the detection and categorization of harassing behaviors online are a consequence of classification. Classification—the foundation of information infrastructures—describes how information is sorted according to recognized patterns to facilitate some improved understanding. However, as Bowker and Star (2000) describe, classification is an inherently human process—and as such, it requires human decision-making about what does and does not belong in a given category. Classification becomes a concern when labeling decisions are made with little consideration of the biases inherent in—and thus, risks associated with—those decisions. Still, technology systems rely on classification as a necessary and robust mechanism for scalability: database systems, user accounts, and social media profiles all require the labeling and categorization of people into a series of digital bits.

Bowker and Star (2000) emphasize that classification systems embody moral choices that reflect greater societal values. They cite 1950's South Africa, in which the Population Registration Act and Group Areas Act required that people be classified by racial group and constrained as to where they could live and work—the precursor for the brutality of South African apartheid. While extreme, we see how classification systems can valorize prevailing or dominant points of view while silencing others (Bowker & Star, 2000). Power, then, is held by those who are creating the labels.

More recently, we see how classification is inherent in everyday technological systems and that classifications are largely socially constructed. For example, in 2014, Facebook revised its gender field from "male" and "female" to allow 56 additional options, including transgender, gender non-binary, and gender questioning identities. As of 2017, Facebook offered 71 total gender categories for users to choose from, including a custom field for users to define themselves (Williams, 2014). From a social justice perspective, this change is welcome and embraced. However, practically, Facebook still needs to serve advertisements as part of its business model, and gender is one criteria used for targeting. As a result, users may not in fact know how their chosen gender categories are subsequently categorized by Facebook's algorithms in order for Facebook to more effectively target advertisements.

In this work, we explore the opportunities and limitations of classification in the domain of online harassment. Using HeartMob as a case study, we suggest that classification can motivate bystanders to provide support, both by demonstrating the breadth and scale of harassment experienced online and by surfacing clear avenues of support. We also show how classification can invalidate harassment experiences, particularly for users whose experiences do not fall within the bounds of a 'typical' harassment experience. As argued by Bowker and Star (2000), classification systems embody the morals and values of their creators—whether in the context of a technical system, such as Facebook's automatic detection of harassing language, or a platform policy, such as Twitter's categorization of what constitutes abusive behavior. When a classification system is created with dominant values and morals in mind, the needs of marginalized users are neglected. To reconcile these tensions, we turn to intersectional feminist theory to further elucidate the limitations of traditional classification systems and to inform potential alternatives.

3.2.4 Intersectional feminist theory and practice

Intersectional feminist theory (Collins, 1990; Hooks, 2000; Ahmed, 2017) holds that various identities (such as race, gender, class, sexuality, religion, disability, and nationality) are inextricably bound in systems of entrenched structural oppression (such as racism, sexism, cissexism, classism, heterosexism, ableism, and colonialism). These systems of oppression "intersect" and cannot be understood independently of each other (Collins, 1990). For those who have many identities that have been historically dominated, the effect of various oppressions is intensified (Collins, 1990; Hooks, 2000). Crenshaw—who coined the term intersectionality—illustrates that under United States law, women are assumed to be white, and Black people are assumed to be men (Crenshaw, 1991). As a result, there are severe gaps in legal protections for women of color, which white women do not experience (Crenshaw, 1991). Thus, the experiences of Black women cannot be understood as additive (e.g., Black + woman), but must instead be understood as intersecting, interdependent, and mutually constitutive (Crenshaw, 1991).

For example, two of HeartMob's co-founders, Courtney Young and Debjani Roy, are women of color who reside in the United States. As a result, they experience oppression both offline and online based on their gender as women and as being African American and South Asian, respectively. Roy's transnational experience as an immigrant from the U.K. to the U.S. has also shaped her experiences. In contrast, three of the present work's authors are women and are white or pass as white. They experience oppression based on their gender identity as women, but not based on their perceived race or immigration status. Intersectionality offers a lens through which to understand how the oppression women of color experience can be both different from and more acute than that of white women. In the context of the current work, intersectionality helps to position individuals' experiences of online harassment as both reflective of and inextricable from systems of structural oppression, such as racism, sexism, cissexism, and so on. This framework allows us to 1) better understand the limitations of current approaches to online harassment and 2) consider user-driven alternatives.

3.2.4.1 HeartMob

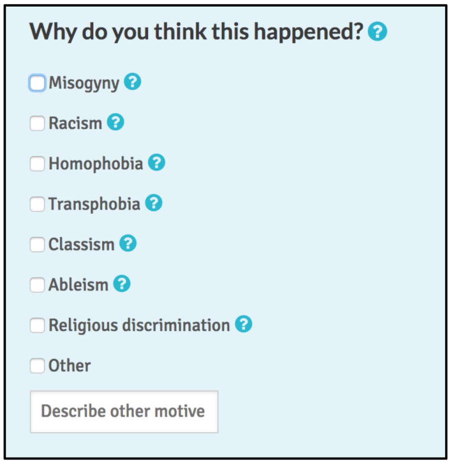
HeartMob (iHeartMob.org), launched in January 2016, is a private online community designed to provide targets of online harassment with access to social and instrumental support. HeartMob was created by leaders of Hollaback!, an advocacy organization dedicated to ending harassment in public spaces. Hollaback! leaders and their colleagues experienced consistent and often severe harassment online as a result of their work. Given their collective expertise in intersectional feminist practice, social movement framing, and bystander intervention to combat harassment in physical spaces (Dimond et al., 2013), the Hollaback! team sought to translate these practices online to support others with similar experiences. Their goals were 1) to understand if, and to what extent, online tools could help create communities of accountability, and 2) to organize people who witness online harassment (i.e., bystanders) to provide support to harassment targets.

HeartMob was designed with an intersectional feminist underpinning: it was built by and for people who are disproportionately affected by the most severe forms of online harassment due to their intersecting oppressions. In late 2014, the team convened a diverse group of journalists, academics, and feminist activists who had all experienced severe online harassment. In the workshop, the second author organized speculative design activities so that participants were directly informing the system's ultimate design. The design of the system went through several design iterations, with continuous feedback from the convening community. As a consequence, the design team learned the ways in which actual targets of harassment wanted to be helped, including supportive (and moderated) messages, a space to document their experiences, and assistance reporting harassment to the platforms on which it occurs. Additionally, participants wanted the option to make their harassment experience public, including a description of the perceived motive.

3.2.4.2 How HeartMob works

The resulting HeartMob system was designed around the specific needs of people who had experienced online harassment. After creating an account, a person experiencing harassment can submit a harassment case, which includes a description of their experience, the type of harassment (e.g., stalking or doxing), the perceived motive (e.g., racist or misogynist harassment), and any screenshots or additional documentation (see Figures 3.1 and 3.2). The user can optionally create a help request and specify the type of help they would like (for example, supportive messages or assistance reporting harassment). The user account, harassment case, and help request are all moderated by trained HeartMob employees. The target can optionally make their harassment case public outside of the HeartMob community to make online harassment more visible.

Users can also apply to be a "HeartMobber," or a bystander who supports targets of harassment by fulfilling their requests for help. HeartMobbers are vetted using a tiered trust system. Level 1 HeartMobbers are only required to verify one social media account but are still moderated. As a result, they are limited to the information they can see about a particular help request until they are accepted as a "Trusted HeartMobber." In order to submit an application to be a "Trusted HeartMobber," users must choose a combination of disclosing additional data and meeting certain criteria based on account duration and approved community participation.



Why do you think this happened?

□ Misogyny

□ Racism

□ Homophobia

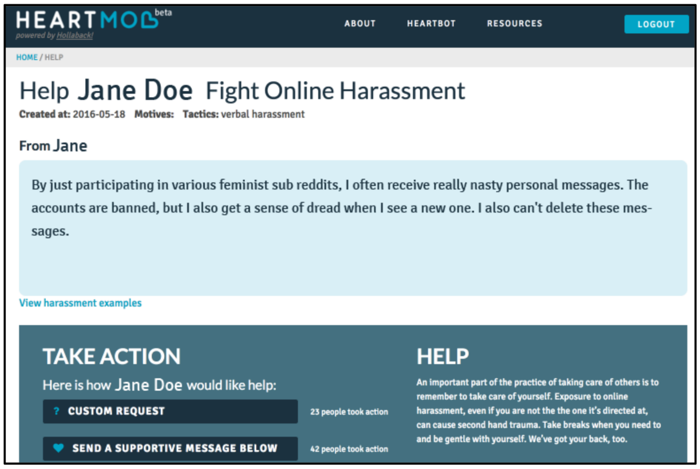
□ Transphobia

□ Classism

□ Ableism

□ Religious discrimination

□ Other



Help Jane Doe Fight Online Harassment

Created at: 2016-05-18 **Motives:** **Tactics:** verbal harassment

From Jane

By just participating in various feminist sub reddits, I often receive really nasty personal messages. The accounts are banned, but I also get a sense of dread when I see a new one. I also can't delete these messages.

TAKE ACTION

Here is how Jane Doe would like help:

HELP

An important part of the practice of taking care of others is to remember to take care of yourself. Exposure to online harassment, even if you are not the one it's directed at, can cause second hand trauma. Take breaks when you need to and be gentle with yourself. We've got your back, too.

Figure 3.1: HeartMob categories for perceived harassment motives.

Figure 3.2: Anonymized HeartMob help request.

Table 3.1: HeartMob supportive actions.

	Percentage of cases with a request for this action	**Median number of actions taken per case**
Supportive Messages; 'Got Backs'	67.6%	41.5; 46.5
Reporting Abuse	70.4%	2.5
Documenting Abuse	46.4%	1
Other	0.7%	9

As of February 2017, there were a total of 1,455 approved HeartMobbers and 98 approved target accounts, producing a total of 86 approved harassment cases and 71 associated help requests. Having a help request is not required, as some targets simply want to share their story. Targets may also request help from individual people instead of (or in addition to) the HeartMobber community. There have been a total of 4,555 actions taken by HeartMobbers on targets' help requests. Supportive messages are the most popular type of help provided on HeartMob, with a median of 41.5 messages and 46.5 "Got Backs" (similar to the Facebook 'Like' button) per help request (see Table 3.1). Although targets most frequently request help reporting abuse to platforms, there are not as many actions taken by HeartMobbers as supportive messages. This is likely due to the various challenges of reporting content to platforms, which will be discussed further in the results.

3.3 Methods

We conducted semi-structured interviews with 18 users of the HeartMob system. As of February 2017, HeartMob had 1,455 total users. Participants were recruited via an email blast sent to current HeartMob users (e.g., anyone who had created an account) beginning in September 2016. All HeartMob users aged 18 and older were invited to participate, whether or not they had personally experienced online harassment. Four email blasts were sent in total, with the last recruitment email sent in January 2017. 25 HeartMob users expressed interested in participating; 18 users ultimately completed an interview. All participants were located in the United States or western Europe, and interviews were conducted in English.

Before the interview, each participant was asked to review and sign an online consent form, which included a short demographic survey. Participants were asked their age (open response); gender (open response); whether they identify as transgender (yes or no); sexual orientation (Heterosexual or straight; Gay or lesbian; Bisexual; or Please specify); and race (White; Hispanic, Latino/a/x, or Spanish origin; Black or African American; Asian; American Indian or Alaska Native; Native Hawaiian or Other Pacific Islander; or Please specify). Participants ranged in age from 28 to 71 years, and the median age was 40.5. 10 of our 18 participants identified as heterosexual; one participant identified as hetero-poly (for polyamorous, or someone who consensually pursues multiple sexual or romantic relationships). Four participants identified as gay or lesbian, one as queer, one as bisexual, and one as pansexual. The high number of non-heterosexual participants in our sample may be partially explained by the increased likelihood of LGB people to experience harassment online (Lenhart et al., 2016).

Table 3.2: Participant demographics (self-identified).

	Age	Gender	Sexual orientation	Race
P1	28	Male	Gay or Lesbian	White
P2	64	Female	Heterosexual	White
P3	43	Male	Heterosexual	White
P4	34	Female	Bisexual	White, Hispanic
P5	36	Male	Gay or Lesbian	White
P6	65	Female	Gay or Lesbian	White
P7	45	*	Gay or Lesbian	White
P8	28	Male/They	Hetero-poly	Italian, Irish
P9	30	Female	Heterosexual	White
P10	42	Female	Heterosexual	White
P11	31	Female	Queer	White
P12	57	Male	Heterosexual	White
P13	51	Female	Heterosexual	White
P14	48	Female	Pansexual	White
P15	71	Male	Heterosexual	White
P16	35	Male	Heterosexual	White
P17	31	Female	Heterosexual	White, Black
P18	39	Female	Heterosexual	Asian

We provided an open text field for participants to identify their gender. Ten participants self-identified as female; 7 participants self-identified as male. One self-identified male participant, P8, included a pronoun preference, and will be identified in the paper as they/them. The remaining participant (P7) chose not to identify a gender*, saying that "gender is a system of oppression, not an identity." Instead, she identified her sex as female. None of our participants identified as transgender. Although transgender and gender non-conforming people are likely to endure severe harassment online due to systemic transphobia, misogyny, and homophobia (Clark, 2015), these experiences remain unrepresented in recent online harassment research (e.g., Lenhart et al., 2016), including in the present study. This is an important limitation that should be explored in future research.

Interviews were conducted between September 2016 and January 2017. Interviews lasted an average of 52 minutes; the longest interview lasted 120 minutes and the shortest 19 minutes. Generally, interviews with participants who had experienced online harassment themselves were longer than interviews with participants who hadn't, due to the personal nature of online harassment experiences and the emotional labor of recounting them. All interviews were conducted and recorded using BlueJeans, a secure online video service. Participants were not compensated. This study was approved by an accredited Institutional Review Board.

Most users had experienced some form of online harassment themselves (n=11; P2, P3, P4, P5, P6, P7, P8, P11, P13, P17, P18), but not all of these participants had submitted a case to HeartMob (some were motivated by their own harassment experiences to join HeartMob as a supporter). We asked these users about their harassment experiences, what support they received (if any), and the process of using HeartMob to seek support (if they had done so). Other participants joined HeartMob purely in a supportive role (n=7; P1, P9, P10, P12, P14, P15, P16); we asked these users about their motivation to join HeartMob, whether they felt a sense of community with other HeartMob users, and about the process of offering support to people experiencing online harassment. We asked all users what the phrase online harassment means to them and about their impressions of the HeartMob site. We sought to elicit specific experiential narratives from our participants through the use of general questions centered on specific emotions (e.g., "Tell me about your most recent experience with online harassment, or about an experience that was particularly difficult" or "Tell me about a time where you felt the support you provided was helpful"). Last, participants were asked what additional support HeartMob and other platforms (for example, social media sites such as Facebook and Twitter) could provide to users like them.

Interview recordings were transcribed by Rev.com. We used an inductive approach to develop codes (Thomas, 2006). One member of the research team individually read through interview transcripts and noted codes by hand. After discussing these initial codes as a research team, we created a more comprehensive list of codes (61 codes in total). Resulting codes were organized around several themes, including but not limited to defining online harassment, impacts of harassment experiences, changes in technology or privacy use, seeking or providing support, and visibility and audience. Two researchers each coded four interview transcripts in a pilot coding process to test and refine the codebook. We coded interviews using Atlas.TI, frequently discussing codes to maintain agreement. Each interview transcript was coded by two members of the research team. Quotations have been lightly edited for readability.

3.4 Position statement

Per feminist methodology, the authors recognize the importance of positioning the research team in relation to the current work and our analysis (Williams & Irani, 2010; Bardzell & Bardzell, 2011). All of the authors have been close with someone who has experienced online harassment, and two of the authors have experienced it personally. Three of the authors are women, and one is a man. All of the authors identify as white. As our sample is also majority white, the absence of experiences from or interpretations by people of color is a significant limitation: because of cultural racism, people of color face different kinds of harassment online than do white people. Further, women of color experience a unique intersection of racist and misogynist online harassment, which is different from the experience of racist harassment or the experience of misogynist harassment. This experiential understanding is notably absent from the current work. Similarly, all authors are cisgender, and none of our interview participants identified as transgender. More research should explore and amplify the online harassment experiences of trans people.

3.5 Results

Results are organized around three major themes: first, for targets, labeling experiences as ‘online harassment’ provides powerful validation of their experiences. Second, for bystanders, labeling abusive behaviors as ‘online harassment’ enables bystanders to grasp the scope of this problem. Third, for online spaces, visibly labeling harassment as unacceptable is critical for surfacing norms and expectations around appropriate user behavior.

3.5.1 Labeling validates targets’ harassment experiences

When a user submits their harassment experience to HeartMob, they select the type of harassment (e.g., stalking or doxing) and the perceived motive (e.g., racist harassment or misogynist harassment) from drop-down lists. HeartMob moderators then individually approve each case of online harassment an individual user submits. Participants felt validated when their experiences were accepted and labeled as 'online harassment.' P5 said:

> "It's the safety net. Right now, the worst that can happen is somebody experiences harassment and they have nowhere to go—that's what's normal in online communities. With HeartMob, if someone says they're experiencing harassment, then at least they get heard… at least they have an opportunity to have other people sympathize with them."

P11 referred to HeartMob as a means of "harm reduction"—she said it "doesn't have the capacity to single-handedly solve the problem, but it makes being online bearable." P9 said that even though she had not experienced harassment online, offering support to others on HeartMob made her feel safer: "Personally, it makes me feel a little safer. Even though I've never used this as someone experiencing harassment, I know that it's there—it's comforting to know there's this network." Similarly, P8 said that supporting other users made him feel like HeartMob is "something I can turn to."

Many participants expressed a preference for support from users who empathized with their unique experiences. Some participants felt that HeartMob's system of labeling enabled them to find other users with similar experiences or shared identities. This finding underscores the importance of understanding harassment through an intersectional lens, as different groups of people experience oppression differently. P13 felt that HeartMob provided immediate access to social support from people who understood the impacts of her specific experiences:

> "I feel like what HeartMob is doing is so instrumental in keeping people from going off the deep end—from feeling alone in this. Most people don't quite understand… how it invades every aspect of your life, basically, when this happens to you. Even my friends—they knew on a daily basis what was going on, and they still couldn't really grasp it."

Other participants felt that more targeted support could help them better prepare and protect themselves in the future. P11, who had several experiences with sustained, high-volume attacks (e.g., dogpiling), said that during her first experience with harassment, she had no idea "how scary it is to see hundreds and hundreds of people wishing death upon you." During subsequent experiences, knowing what she could expect helped P11 to “rally her troops” and preemptively seek support for potential harassment:

> “I was able to tell the folks around me, ‘Hey, this is gonna be really rough, so I'm gonna need you to send me some love. If you see an article posted on someone's Facebook and you see nasty comments, I'm gonna need you to come to my defense.’ People really came to my defense, which was incredible to me… I didn't have that [the first time], because I had no idea what was coming.”

P13, whose former partner created a number of defamatory websites that had disrupted her personal and professional life for years, felt isolated from her friends and family and had a difficult time accessing relevant information and legal resources. After joining HeartMob to support other targets of online harassment, P13 suggested that a system for categorizing submitted cases according to specific harassment behaviors and their impacts could help connect isolated users with the support they need most: “I think the person would feel a little more connected. I feel like we’re all in silos.”

3.5.1.1 Classification privileges dominant experiences

Participants like P13, who was not able to use HeartMob to identify others who shared her circumstances, were likely to minimize the impact of their own experiences. Often, these participants made comparisons to harassment experiences they had observed or perceived others to be experiencing. P3, a 43-year-old white man, had lost several employment opportunities due to defamatory information posted about him online by a former associate—and had spent thousands of dollars trying to expunge the defamatory information from his online record. P3 joined HeartMob specifically to access other users who could directly empathize with his experiences: “When I registered with HeartMob, my only intention was to get some kind of support network... I hoped to identify with someone that’s dealing with something similar to me. I don’t think there are a lot of people out there that experience what I experienced.”

P3 said that first joining HeartMob had been difficult, in that he didn’t know how to best categorize his experience using the system’s available labeling tools. Ultimately, P3 felt that his inability to use the HeartMob system to identify users experiencing similar abuse signaled that his experiences might not ‘count’ as online harassment:

> "Trying to find the right checkbox to categorize yourself is tough sometimes… when someone thinks of online harassment, they don't think of what I've been going through. I don't even think people would define my case as online harassment. I know there are people out there that just don't care what I'm going through."

P3 ultimately had difficulty soliciting the unique support he needed from the HeartMob community, and he suggested—based on other cases he had seen on HeartMob—that other, more marginalized users were more deserving of the community's support:

> "The lesbian and gay community, how they are discriminated against online and harassed—that's more important than my situation. My case probably seems very small and insignificant, considering. I don't think the world would feel sorry for someone like me."

For P8, whose middle school classmates had once impersonated him online to harass a favorite teacher, reading other users' harassment cases on HeartMob made him doubt the severity of his own experiences: "I've never experienced anything where I have felt threatened for my safety… nothing like what I see on HeartMob." P13 similarly minimized her own harassment experiences. When P13 ended a romantic relationship, her former partner created a defamatory website suggesting P13 had participated in prostitution. He included P13's contact information and professional history, and circulated the website to over 300 of P13's friends, family members, colleagues, and clients. P13 had been working with local and federal law enforcement and with Google for more than four years to have the defamatory sites removed. Still, P13 said: "At one point there were 14 websites. I feel like 14 is a lot—but as I read other people's stories [on HeartMob], I realize it could be a lot worse."

3.5.1.2 Labeling reflects community norms

Some participants did not see their own identities reflected in the cases ultimately posted to HeartMob (i.e., approved by moderators), and as a consequence, they questioned whether or not their experiences belonged—and by extension, whether or not they belonged—on the site. When asked whether she had sought support on HeartMob for her own harassment experiences, P2, who is 64 years old, said she hadn't considered it because the interactions she had witnessed on the site left her with the impression that most users were significantly younger: “It didn’t occur to me… I’m there to provide support. The bulk of people I [am supporting] are much younger.”

Some participants felt that system labels on HeartMob privileged certain perspectives over others. P7 had applied to become a trusted HeartMobber, but had been denied—which P7 suspected was a direct result of writing publicly about her controversial views on gender, which others had characterized as abusive. P7 felt that she had been unfairly characterized by HeartMob, and as a result, she felt she would not be welcome to solicit support on HeartMob for her own harassment experiences:

> “I think that HeartMob comes from this leftist liberal mindset. When people who have political views that are supported by a community engage in that type of behavior, they're praised. When people like me—who have unpopular views—engage in that type of behavior, we're accused of abuse. I would hazard a guess that if someone like me went to [HeartMob] and was like, ‘I'm really being harassed,’ they probably wouldn't help me, honestly.”

Participants also sought ways to more easily locate users who shared their social identities: P5 joined HeartMob specifically hoping to support other LGBT users experiencing harassment online, but he could not locate experiences similar to his own experiences of harassment and exclusion on Wikipedia: “I could browse other people’s stories, and there were no LGBT stories in the queue. I thought I might have wanted to put my story in an LGBT queue, but there was no such thing at the time.” Most participants felt that support services for people experiencing online harassment should welcome a diversity of users and perspectives. Said P7: “I think they should make it very clear that they are agnostic about who they support, and that it doesn't matter what you believe. If you’ve been abused, they'll support you. I think that would be great.” However, this perspective may be at odds with HeartMob’s original design goals, which prioritize marginalized users and intentionally script these values into the design of the system itself.

3.5.1.3 System labeling invalidates harassment impacts

Participants also felt invalidated when their experiences were outright rejected by the system, or otherwise labeled as not harassment. This was particularly salient for social media users, as most major social media companies rely on scripted responses that do not acknowledge individual experiences or the impacts of being harassed. P17, who endured large volumes of harassment as a result of her work as a writer, felt that the labor required to report ongoing harassment was "completely useless." She continued: "There's really no point in reporting stuff on social media. Last time, I spent several hours going through and reporting tweets. It felt like maybe less than 10% were considered threatening—and either they had their account indefinitely suspended, or just suspended until they took the tweet down." Reporting was particularly frustrating for users if reported content was ultimately found not to be in violation of existing policies, which many participants felt was a frequent occurrence. P17 went on to say:

> "What I think was really frustrating was the level of what people could say and not be considered a violation of Twitter or Facebook policies. That was actually really scary to me—if they're just like, 'You should shut up and keep your legs together, whore,' that's not a violation because they're not actually threatening me. It's really complicated and frustrating, and it makes me not interested in using those platforms.

Even when social media platforms removed abusive content, participants felt the process by which a verdict was reached (or a sanction determined) was obscure. After P18 and her colleagues had publicly launched a new product, some users threatened their physical workplace. When P18 was sent an image on Twitter of a man pointing a sniper from a rooftop, she reported the tweet: "We did ask Twitter to take that down, and they did—but I don't know what they did with the person who posted it. I don't remember what happened with that." The challenges of reporting abuse directly to the platforms where it occurs were similarly frustrating for participants who wished to support harassment targets, which may explain why so few HeartMobbers provide this type of support—despite it being the type of help most frequently requested by targets (See Table 1). P9 had tried to help a target report abuse, but the content was no longer available: "I went online to the link the person had posted, but all of the posts were already gone. I guess somebody had—somebody must've already reported it. You could see that two people had responded to the request, but that was kind of the end of it." This gap between system labels and user experiences led many participants to wonder whether targets were receiving the support they requested, or whether they were doing enough to support them. P14 said: "There was somebody who had some really significant, terrible comments. I did worry that my words of hope and encouragement in that situation weren't enough. I felt I could have been far more helpful."

3.5.2 **Labeling reflects community norms**

We find that for bystanders, labeling the variety of abusive experiences enabled by technology as 'online harassment' helps them understand the breadth and impact of this problem. For P14, participating as a HeartMobber changed her perceptions about the severity of online harassment. P14 felt that providing support on HeartMob helped her better understand the breadth of online harassment experiences:

> "In the work that I do, I help people understand how society plays a role in the violent culture that we have—but I've never had too much opportunity to actually see the evidence on the internet. I knew it was there. I talk about it, present about it, but actually seeing the horrific things that people are seeing and doing to others online really brought that to a whole different place for me."

P11 suggested that visible disclosures of harassment experiences, like the collection of cases available on HeartMob, could encourage targets to seek support:

> "The second someone's vulnerable about their experience with violence online, it creates this sense of community amongst others who've also had that experience. It's like, oh my gosh, you too—we're some of the only people talking about this, so we have to have each other's back."

Still, participants worried that targets may make themselves vulnerable to additional abuse by publicly disclosing their experiences. P9 felt that even a protected space like HeartMob could jeopardize targets' ultimate safety and comfort online: "If somebody hacks in, it's so much more traumatizing if you get harassed from within that space."

HeartMob also exposed users to the diversity of abusive behaviors used to target people online. P17 expressed that online harassment could include a range of potential experiences: “It’s anything—email, direct messages, in the comments. It’s the doxing of people’s information. It’s sustained threats, and sometimes just little ones here and there. It can happen either once or over a sustained period of time.” P3, who had experienced an atypical form of online harassment, acknowledged that it was difficult for others to empathize with his experiences: “I think until you live this nightmare I've been living with, you just don’t know.”

P11 suggested that the problem of online harassment is particularly insidious because bystanders—particularly those who have not experienced harassment themselves—feel powerless to help. By classifying different harassment experiences and providing specific, labeled ways in which to provide support, P11 felt HeartMob made it easier for bystanders to offer support: “HeartMob is a brilliant way of addressing a problem that I think immobilizes most people, because it seems so big and daunting—so they don’t do anything at all.” P1, who worried that his family members and friends did not consider online harassment to be a ‘real’ problem, felt his participation in HeartMob was a good first step toward speaking out against online harassment in more public spaces: “I think it's very good participating in these cases while not exposing yourself to abusers in public spaces. I think of HeartMob as the stepping stone to participating in the public spaces.”

3.5.3 Labeling helps define responses to harassment

Last, we find that visible resistance to harassment in online spaces is important not only for targets, but for surfacing norms about what is and is not acceptable behavior. P11 said that online harassment is not taken seriously as a problem, and that in particular, people do not recognize the full breadth of the problem:

> “When I think of the phrase online harassment, I think about death by a thousand cuts. I think about how we either don't take it seriously as a society, and we think it's just the internet and turn it off, or we view individual tweets or emails or comments in isolation—we don't view the breadth of it, recognize the avalanche.”

3.5.3.1 Resisting normalization

For many participants, seeing other users support harassment targets on HeartMob provided visual evidence that demonstrated the full scale of user resistance to the problem of online harassment. This community opposition to online abuse was not always apparent on other platforms. P9 noted that online harassment can be particularly isolating when targets are exposed to significant volumes of abuse, but receive comparably limited support:

> "It's something that's very isolating, because it can make you feel—especially if there are multiple people doing the harassing—like everybody would be against you… like they're representing society."

P9 went on to say that harassment perpetrated by anonymous or pseudonymous users can seem representative of society at large: "Because it's anonymous, it can give you this sense that they're representative. It makes you lose faith in humanity. There are just so many people devoting their time to bringing other people down."

Participants who regularly experienced harassment and abuse online admitted to feeling like their experiences had become normative. P7 said that she had become indifferent to online harassment over time, due to the volume of threats she had experienced:

> "It's like exposure therapy. After time, it's alright. You're able to put boundaries around what's a real threat, what isn't actually a real threat—even though it might hurt my feelings, and even though I might feel my body react or my heart raise."

P17 agreed: "It's annoying, because I'm getting used to it. It's just becoming the norm for how our society conducts itself, how people discuss things and interact." P11, a writer, said that after witnessing so many others experience online abuse, her first experience felt familiar: "There was this familiarity… it was like I had been initiated into the experience." Her friends were unsurprised or even apathetic, P11 said: "I was met with a lot of like, yeah, welcome to being a woman online. Welcome to engaging with American media. Welcome to the club. There was this apathetic, cynical response that I found very discouraging."

Participants felt that more visible resistance to harassing behaviors online would discourage these types of experiences from becoming normalized. Participants also worried that the normalization of harassing behaviors online impacted other users. P17 worried about the impact of her public harassment on other users: "I definitely get messages from people who are like, 'I want to share my story, but I see what you go through.'"

3.5.3.2 Reclaiming space through labeling

Traditionally, social movements have taken to public spaces, such as streets, to reclaim space as an act of resistance. Similarly, participants sought to reclaim online spaces, where abusive behaviors were perceived to occur frequently. For participants who had experienced harassment, social media site use became undesirable or even frightening. P17, who had taken a hiatus from social media sites as a way to avoid ongoing harassment as a result of her work as a writer, said: “The work that I do, I have to use [social media]. It is not a tool that is fun for me anymore. It is a tool for work.” P17 discussed the need for a tool to mitigate high volumes of public abuse, “to break up the monotony of the hatred.” Specifically, P17 felt overwhelmed by the sheer amount of abuse visible in her social media notifications:

> “You know that scene [in Harry Potter] where Harry uses his Patronus for the first time, and it knocks out all the Dementors? It’s this white light that pushes everybody back. I wish there were—I don’t know the social media equivalent of that. I need a Patronus right in the moment to just push everybody back.”

Participants perceived online spaces as being overrun by harassment and abuse, with little visible resistance. P13 said: “The internet has become the town square where people are taken to be embarrassed and punished. The only difference is the whole world can see it, not just the town. Then it stays forever—it doesn’t go away.” P7 felt that it was important to define the line between abuse and disagreement on social media sites, particularly in a divisive political climate:

> “In order to have a free and fair and just society, all people need to feel empowered to speak. It gets tricky on social media, when people are designating certain things as abusive speech that are actually political disagreement. For example, I'm a Democrat. Someone who’s a hardcore Trump supporter, I might not like them, I might think that they're stupid, but they have every right to their opinion, and it's not hate speech for them to post that they think Trump should build a wall. I might think it's racist, I might think it's wrong, but it should be permissible for people to say that.”

P12 felt that it was important to emphasize civility online: “We've got to start taking back the internet. It's got to come back to what it was—a place of information, a place of sharing ideas. If I go on a website and I don't like what they're talking about, I can leave. I don't have to attack them." P13 felt that seeing greater resistance to harassment in online spaces would encourage her to express visible support for targets: "If I go online to support a person, and then I look at the other messages to that person—I think, 'Oh, that's neat. There are other people out there doing it, too.'" One way to reclaim online spaces is through making norm violations visible and labeling them as harassment. Many participants reflected on the relationship between a perceived increase in harassing behaviors online and the current political climate at the time of data collection. Several participants (n=5; P13, P17, P2, P15, and P10) specifically mentioned Donald Trump as having participated in or incited online harassment, including his criticism of Chuck Jones on popular social network site Twitter. Jones, the president of United Steelworkers 1999 (a labor union), suffered substantial harassment following Trump's tweet (Shear, 2016). P17 felt President Trump's comportment would normalize similar behaviors:

> "I just had this moment of, 'Oh my god; this is not going to end.' I had this realization… it's happening from the top, and we now have a president who condones this. That terrifies me. There are millions of Twitter users who see that this behavior he's doing is okay—that this is how you conduct yourselves when talking to people who disagree with you."

P5 felt that social media platforms could also benefit from exposure to a diverse corpus of documented harassment experiences. P5 was concerned that internet companies may not be equipped to fully understand the experiences of their diverse users: "For example, the Wikimedia Foundation—they're based in San Francisco, but Wikipedia gets international harassment problems. If harassment happens in, say, India, white people in San Francisco don't really know what it's like to be a teenage gay boy in India." P5 felt that companies could be doing more to partner with activist platforms or support organizations where diverse harassment experiences can be visibly aggregated. This desire to reclaim civility in online spaces suggests that labeling behavior as abusive and unwelcome plays an important role in defining normative responses to harassment within a given community.

3.6 Discussion

Classification—whether instantiated by technical systems, platform policies, or users themselves—plays a critical role in validating and supporting online harassment experiences, as well as enabling bystanders to intervene. We discuss our results through the lens of Bowker and Star's classification theories and Becker's labeling theory of deviant behavior. We discuss the implications of our results for current approaches to labeling and classifying online abuse, such as platform reporting tools and machine learning techniques for the automatic identification of harassing language. Finally, we present alternatives for representing diverse experiences in online systems.

3.6.1 Surfacing social norms through visible classification

Many of our participants discussed feeling uneasy about or apathetic toward the ways in which abusive behaviors are seemingly becoming normalized online. Participants who were harassed publicly (e.g., dogpiling on social media sites or defamation in the media) often desired more public demonstrations of support than the HeartMob system could realistically provide. This result suggests that visibly labeling harassing behaviors as inappropriate—whether by users or by the system itself—can make opaque, system-driven classification systems more visible, and consequently may help users to identify and define the boundaries of appropriateness in online spaces.

3.6.1.1 Visible labels create powerful descriptive norms

Empirical evidence suggests that injunctive norms—expectations for how you should behave, such as those articulated in a platform policy—are often less powerful than descriptive norms (how most others behave) in encouraging behavior change (Cialdini, 2007; Goldstein et al., 2008). Cialdini (2007) argues that descriptive norms offer "an information-processing advantage," in that by understanding how most people behave in a given situation, individuals can more quickly decide how to behave themselves. Similarly, Marwick (2012) argues that social media users "monitor each other by consuming user-generated content, and in doing so formulate a view of what is normal, accepted, or unaccepted in the community." Users who are uncertain about how to behave will adapt to visible descriptive norms, particularly in cue-sparse online environments (Walther, 2002).

Cheng et al. (2017) found that perpetrators of online harassment were not, contrary to popular narratives, anti-social actors dedicated to violating norms of civility. Rather, people could be primed to harass others online in an experimental setting when earlier instances of harassing behavior were made visible in a comment thread. Where harassment is visible—and any sanctioning of that harassment is not—then the descriptive norm could easily become that harassment is an appropriate activity. If a new Twitter user is exposed to high volumes of harassing content on the platform with little visible resistance (e.g., platform- or user-issued sanctions), for example, the user may determine that harassing behavior is appropriate on Twitter. Our results suggest not only that visible sanctions serve as important validation for harassment targets, but also that public demonstrations of support may help other users determine what behaviors are and are not appropriate in a given space.

3.6.1.2 Visible classification penetrates the "fog of audience"

Computer-mediated communication can be cue-sparse and persistent, making it difficult for users to assess who has seen what and when (Blackwell et al., 2016; Schoenebeck et al., 2016). In discussing the implications of using technology to engage in social surveillance, Marwick (2012) argues that networked online platforms give users an ambient awareness of others, but obscure critical details such as social context and differences in power—a phenomenon which Lampe (2014) refers to as the "fog of audience." A digital audience is "large, unknown and distant" (Palen & Dourish, 2003), and users must continually negotiate their own privacy and impression management practices to satisfy both their own expectations and those of a functionally invisible audience (Palen & Dourish, 2003; Vitak et al., 2015; Litt & Hargittai, 2016). Any one user cannot reliably determine the experiences or expectations of all others in the network, making it difficult for users to assess the norms of a given community.

The difficulty in detecting the norms around harassment on a social media platform is further complicated by policy-driven classifications, which are not typically made visible to users. Many online platforms rely on policies to enforce formal (i.e., codified) norms for what is and is not appropriate behavior when using their services. Although the policies themselves are accessible to users, how and why those policies are actually enforced is more opaque—and any one user is typically unaware of how other users expect policies to be applied, or how they experience policy enforcement. In the absence of transparency, users are left to decide for themselves why platforms make certain choices, and may consequently ascribe values to a system that its creators did not intend. For example, in June 2017, ProPublica revealed that U.S. congressman Clay Higgins' Facebook post calling for the slaughter of "radicalized" Muslims was not removed by the site's content moderators (Angwin & Grassegger, 2017). Conversely, a post by Black Lives Matter activist Didi Delgado—"All white people are racist. Start from this reference point, or you've already failed"—was removed, and resulted in a week-long account suspension (Angwin & Grassegger, 2017). Although it is unlikely Facebook's content moderators made any direct comparisons between these two posts when deciding how best to enforce their policies, the obscurity of Facebook's classification system leaves users to draw their own conclusions about what is or is not appropriate on Facebook and why.

Importantly, Bowker and Star (2000) argue that classification systems—though ubiquitous in everyday lives—are typically invisible, and thus people remain largely unaware of the social and moral order they create. Classification systems are typically made visible only "when they break down or become objects of contention" (Bowker & Star, 2000). As documented by ProPublica (Angwin & Grassegger, 2017), Facebook content moderators are trained to recognize hate speech as curses, slurs, calls for violence, or other attacks directed at "protected categories," which are rooted largely in Western legal definitions (i.e., sex, race, ethnicity, sexual orientation, gender identity, religious affiliation, national origin, and serious disability or disease). However, factors such as social class, age, and occupation are not protected under U.S. law, nor are categorical factors such as religions or countries (in other words: attacking an individual's religious affiliation is not allowed, but attacking a specific religion is). As a result, ProPublica argues, Facebook's algorithm is designed to "defend all races and genders equally"—an approach which Danielle Citron, quoted within, argues will "protect the people who least need it and take it away from those who really need it" (Angwin & Grassegger, 2017). Though Facebook was the focus of this particular report, most online platforms—including Twitter, Craigslist, and others—must present publicly available policies (e.g., community standards) while also engaging in often "invisible" classifications to effectively enforce these policies at scale. This obscurity not only contributes to uncertainty among users about what is and is not considered acceptable behavior in online spaces, but also creates distrust about which values these technologies privilege.

3.6.2 Classification reifies oppression

Classification systems impact the ways in which social norms are created and enforced—for example, the medical profession's classification of homosexuality as an "illness" during the nineteenth century led to stigmatization which persists even today (Weeks, 1999). It is important to note the relationship between social deviance and social oppression: what is considered "deviant" in a given society is defined by dominant social forces, and thus deviant labels may unjustly malign members of marginalized groups. The importance of social labeling in the emergence and persistence of social norms (Weeks, 1999) is critical to understanding not only societal perceptions of behaviors, but also the ways in which labels may be leveraged in the oppression of non-dominant groups.

As Bowker and Star (2000) emphasize, classification systems emphasize the concerns of dominant groups, and are often created specifically to impose dominant norms upon oppressed persons. Marxist social conflict theory (Marx & Engels, 1848/1967) similarly defines deviant behaviors as those which conflict with the goals of social institutions and the ruling class. Rooted in the recognition of structural differences in power and social class within capitalist societies, conflict theory (Marx & Engels, 1848/1967) asserts that more powerful social groups are motivated to retain their power over oppressed groups, and as such, they assert that power through the application of laws and other classifications for behavior designed to oppress less powerful groups.

A promising direction for addressing online harassment is through the automated detection of abusive words and phrases (e.g., Yin et al., 2009; Chandrasekhara et al., 2017; Wulczyn et al., 2017); however, our findings show that, like in other cases where descriptive norms are being constantly negotiated, accounting for fluid social nuance may prove challenging for automated approaches. Similarly, Crawford and Gillespie (2016) surface critical limitations of technical tools that enable users to flag content: for example, Facebook's removal of a photograph of two men kissing following its flagging by several users as 'graphic sexual content.' Although Facebook ultimately reinstated the image and apologized, this incident illustrates the limitations of sociotechnical classification systems for labeling non-normative—or non-dominant—behavior. Crawford and Gillespie (2016) argue that flagged content is not a proxy for non-normative or deviant behavior, but instead represents complex negotiations between platforms, individual users, and broader regulatory forces. As such, automatic classifiers for detecting harassing language or behaviors are far from a simple solution, and indeed could exacerbate existing structural inequities.

3.6.2.1 System classifications exclude outsiders

Bowker and Star (2000) describe classification systems as “artifacts embodying moral and aesthetic choices that in turn craft people's identities, aspirations, and dignity.” Our participants reported that they sometimes felt their harassment experiences did not fall within ‘typical’ expectations of what online harassment looks like, or even who is harassed online, especially those who were not able to accurately categorize their experience when submitting their case using HeartMob’s checkbox system. This is reminiscent of Becker’s (1963) labeling theory, which posits that socially applied labels can affect individuals’ self-conceptions and behaviors, making those labeled as ‘deviant’ more likely to see themselves as outsiders.

In the language of Becker (1963), our participants saw themselves as outsiders as a result of HeartMob’s classification of their experiences. In this way, the HeartMob system itself acts as a boundary object (Star & Griesemer, 1989), moving the enactment of power from an automated approach (e.g., Twitter and Facebook’s automatic detection of and response to potentially abusive behavior) to human moderators. HeartMob moderators must still make classification decisions about what experiences should and should not be prioritized in the HeartMob community, which inevitably leads to the exclusion of some users who may belong to more privileged groups, or whose experiences do not fit within typical categories.

These concerns resurface when corporations—who are managing policies at an enormous scale and who are accountable to their shareholders—must create content moderation policies that reflect the values of the company, yet are enforceable at scale. To return to the Facebook example, ProPublica’s article (Angwin & Grassegger, 2017) accuses the company of favoring “elites and governments over grassroots activists and racial minorities,” based on the examples given (a U.S. congressman and a Black Lives Matter activist). From a social conflict theory perspective, the values and worldviews embedded in a company—driven by its leaders, shareholders, and employees—are likely to be reflected in its formal policies, which classify particular behaviors as appropriate or inappropriate. Instead, a more effective system for classifying appropriate behavior may be one that is co-governed between platforms and their users, which would allow for the introduction of additional social nuance in platform policies while fostering a greater sense of accountability among users.

3.6.3 Implications for social change

There cannot be a neutral or value-free approach to harassment categorization (Bowker & Star, 2000). Intersectional feminist theory expands theories of power and oppression by linking different categorical identities—such as gender and race—to systems of structural oppression, such as sexism and racism. We argue that social media platforms and others working to prevent, detect, or manage online harassment must consider power and oppression when creating classification systems, including reporting tools, moderator guidelines, and platform policies.

First, platforms should make visible and disclose the categories, criteria, and process by which harassment is categorized. By rejecting or accepting incidents of online harassment without full disclosure around the values by which such decisions are made, targets may feel invalidated, causing additional harm and potentially affecting their future technology behaviors (e.g., chilling effects). Our data shows that when users receive scripted responses from platforms that do not directly address their specific experiences—or worse, when users receive responses that indicate there has been no violation of a specific site policy, as many of our participants did—targets experience increased social isolation and may subsequently minimize the impacts of their harassment experiences. Thus, system validation for harassment experiences through labeling is critical for targets' emotional health and recovery.

3.6.3.1 Addressing online harassment by centering vulnerable users

Ultimately, our results suggest a need for more democratic, user-driven processes in the generation of values that underpin technology systems. One movement towards this goal can be observed in the platform cooperativism and worker cooperative movements, which seek to turn users and employees into cooperative owners of—and participants in—such systems. Future research should further explore innovative democratic practices around online harassment management and support, particularly efforts driven by users.

It is important to emphasize that platform-driven design often privileges the experiences and concerns of socially dominant groups. Of 11,445 developers in the United States surveyed by Stack Overflow in 2017, 85.5% were men, a majority of whom were also white (Collins, 2017). According to the United States Bureau of Labor (2015), women held 25% of computing-related occupations in the United States in 2015—a percentage that has been steadily declining since 1991, when it reached a high of 36 percent. Nearly two-thirds of women who held computing occupations in 2015 were white: only 5% were held by Asian women, 3% by Black or African American women, and 1% by Latina or Hispanic women (U.S. Department of Labor, 2015). Thus, white men—who in the United States possess the most structural power (Marx & Engels, 1967)—are largely responsible for the ideation and development of policies, moderation guidelines, reporting tools, and other technologies aimed at preventing or managing harassment online: a problem disproportionately experienced by marginalized people (Duggan et al., 2014; Lenhart et al., 2016; Duggan & Smith, 2017).

In applying intersectional feminist theory, we argue that abuse mitigation practices must ultimately protect and be informed by those who are most vulnerable, or the people who historically experience structural oppression. Centering the oppressed in the ideation and development of technology creates stronger objectivity (Harding, 1992) in the categorization of online harassment. A "bottom-up" approach to system design allows us to start from the experiences of those who have traditionally been left out of the production of knowledge, ultimately resulting in technologies that better address the needs of all users. Best addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms.

3.7 Conclusion

Harassment and abuse remain a pernicious problem for modern online communities. Through interviews with 18 users of HeartMob, a system designed by and for targets of online harassment, we find that classification is critical in validating and supporting harassment experiences. We also find that labeling abusive behaviors as 'online harassment' enables bystanders to grasp the true scope of this problem, and that visibly labeling harassment as inappropriate is critical for surfacing community norms and expectations for appropriate behavior. We discuss these results through the lens of Bowker and Star's classification theories and Becker's labeling theory of deviant behavior, and we caution that visible classification systems can also marginalize users whose harassment experiences are not typical, or whose experiences are not accounted for in the system's development. We surface significant challenges for better incorporating social context into existing technical systems, which often fail to address structural power imbalances perpetuated by automated labeling and classification. Similarly, platform policies and reporting tools are designed for a seemingly homogenous userbase and do not account for individual experiences and systems of social oppression. Finally, informed by intersectional feminist theory, we argue that fully addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms. Centering the oppressed in the ideation and development of technology ultimately results in technologies that better address the needs of all users.

Acknowledgements

This work was partially supported by the Knight Foundation. We thank Emma Gardiner for assistance with data analysis and Vidhya Aravind, Renée Cymry, Jennifer Rubin, and Sarah Vieweg for feedback on drafts.

Chapter 4 When Online Harassment Is Perceived as Justified[1]

Most models of criminal justice seek to identify and punish offenders. However, these models break down in online environments, where offenders can hide behind anonymity and lagging legal systems. As a result, people turn to their own moral codes to sanction perceived offenses. Unfortunately, this vigilante justice is motivated by retribution, often resulting in personal attacks, public shaming, and doxing—behaviors known as online harassment.

We conducted two online experiments (n=160; n=432) to test the relationship between retribution and the perception of online harassment as appropriate, justified, and deserved. Study 1 tested attitudes about online harassment when directed toward a woman who has stolen from an elderly couple. Study 2 tested the effects of social conformity and bystander intervention. We find that people believe online harassment is more deserved and more justified—but not more appropriate—when the target has committed some offense. Promisingly, we find that exposure to a bystander intervention reduces this perception. We discuss alternative approaches and designs for responding to harassment online.

4.1 Introduction

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users. This work is motivated by recent examples of harassment in online contexts that, although broadly viewed as harmful, are considered by some as justifiable responses to perceived social norm violations—a controversial form of social sanctioning. This "retributive harassment" can take many forms: high-profile examples include the 2013 public shaming of public relations executive Justine Sacco, the 2015 release of 40 million Ashley Madison users' personal and financial information, or the 2017 doxing of people who attended a white supremacist rally in Charlottesville, Virginia. Retributive

[1] Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When Online Harassment is Perceived as Justified. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

harassment is especially widespread on social media sites such as Facebook and Twitter; however, why it happens and how to prevent it remain unknown.

Historically, abusive behavior online has been relegated to fringe cases—"narcissists, psychopaths, and sadists" (Buckels et al., 2014) who are either exceptions themselves, or inhabit atypical parts of the internet. Today, however, almost half of adult internet users in the U.S. have personally experienced online harassment, and a majority of users have witnessed others being harassed online (Duggan et al., 2014; Lenhart et al., 2016; Duggan, 2017; Rainie et al., 2017). Although policies, reporting tools, and moderation strategies are improving (e.g., Perez, 2017), most online platforms have failed to effectively curb harassing behaviors (Lenhart et al., 2016; Rainie et al., 2017), and internet users and experts alike believe the problem is only getting worse (Rainie et al., 2017).

This research aims to understand online harassment using a retributive justice framework. Retributive justice refers to a theory of punishment in which individuals who knowingly commit an act deemed to be morally wrong receive a proportional punishment for their misdeeds, sometimes referred to as "an eye for an eye" (Carlsmith & Darley, 2008; Walen, 2015). Retributive justice relies upon the assumption that everyday citizens possess intuitive judgments of "deservingness" that accurately and consistently express the degree of moral wrongdoing of others' acts. The integration of theories about justice and punishment with existing knowledge about social deviance and sanctioning has the potential to transform our current understanding of misbehavior in online spaces—in particular, when an instance of online harassment is perceived to be justified.

We conducted two online experiments to test the relationship between retributive justice and the perception of online harassment as justified or deserved. The first experiment tested whether exposure to a retributive prime—i.e., that the person being harassed had committed a crime—increases the belief that harassment is justified, deserved, or appropriate. The second experiment tested the effects of social influence on online harassment; specifically, whether conformity increases the belief that harassment is justified, deserved, or appropriate, and whether or not the presence of a bystander intervention would reduce these beliefs.

Investigating the relationship between orientations of justice and the perception of harassing behaviors online is an important step in better understanding what may motivate users to perpetrate online harassment—as well as what motivates the decisions of moderators and bystanders, who may choose to take action (e.g., flagging or reporting) only against users whose actions they do not perceive to be justified. Ultimately, this research could generate a new understanding of social sanctioning online, influencing the design of technologies that support alternatives to retribution.

4.2 Related work

Online harassment refers to a wide variety of abusive behaviors online, 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 visible humiliation intended to damage a person's reputation). These tactics are often employed concurrently, particularly when many individuals, acting collectively, target just one individual (sometimes referred to as "dogpiling").

4.2.1 Regulating online behavior

The first wave of Internet regulation, emerging in the 1980s (Rheingold, 1993), involved establishing norms for good behavior and sometimes assigning community members special privileges (e.g., admins and moderators) to enforce those norms (Dibbell, 1998; Kraut et al., 1996; Kraut & Resnick, 2012; Lampe & Johnston, 2005; Lampe et al., 2010). Regulation was also supported through moderation tools, such as reporting, flagging, and editorial rights (Matias et al., 2015; Pater et al., 2016). A second wave introduced crowd-sourced approaches to regulation, such as the decentralized approaches used by Slashdot and Digg (Lampe & Resnick, 2004; Poor, 2005). These community moderation approaches have been effective in small online communities, such as LinuxChix, and sometimes in larger communities, such as Wikipedia (Bryant et al., 2005; Kraut & Resnick, 2012; Panciera et al., 2009); however, the size and scope of many online interactions have now outgrown normative regulation. The WELL had three thousand users in 1988 (Rheingold, 1993); Twitter had 300 million monthly active users in July 2017 (Tsukayama, 2017).

Many emerging self-governance techniques in online communities, such as encouraging communities to establish their own rules (Matias, 2017), cannot be implemented at scale. A more recent wave of regulation uses natural language processing and machine learning techniques to generate classifiers for detecting abusive language (Chandrasekharan et al., 2017; Hosseini et al., 2017; Wulczyn et al., 2017; Yin et al., 2009). Though automated approaches have improved dramatically, they are subject to false positives and true negatives, with some harmful content eluding detection while other innocuous content is sanctioned (Hosseini et al., 2017). Furthermore, automatic detection efforts are relatively easy to bypass through subtle modification of language (Hosseini et al., 2017).

This work focuses on what we consider a fourth wave of regulation: everyday users enacting regulation by taking justice into their own hands. Many features of online interactions, such as anonymity, ephemerality, and persistence, are linked with impunity and freedom from "being held accountable for inappropriate online behaviour" (Diakopoulos & Naaman, 2011; Hardaker, 2010). Because offenders face little accountability for their actions online—and because legal systems are often unavailable or ineffective in online contexts—users have turned to forms of "vigilante justice" to enact punishments (Ronson, 2015).

Certain affordances of online platforms, such as persistence, visibility, and broadcastability, may further enable this particular form of justice-seeking. On social media sites, users can easily capture and circulate content, even if the original author later deletes the post. Archived profile histories allow users to make character assessments quickly. When combined with the lack of affective cues in online contexts (Walther, 1996), people's emotional arousal when faced with perceived injustice may lead them to rush to judgment and "fill in the blanks" about others they encounter online. Users can broadcast their desire for justice to wide audiences, and they can easily direct specific sanction requests to an offender's employer, family members, or other visible ties. Further, technological features such as likes, retweets and upvotes promote perceptions of endorsement—known as social proof—that can in turn lead to herd-like behaviors (Schultz et al., 2007; Steele et al., 2002). This has led to extreme and often disproportionate punishments for perceived offenses committed or circulated online, such as public shaming, physical threats, job termination, and sustained social isolation (Ronson, 2015; Sydell, 2017). Just as critically, these vigilante punishments can degrade civic discourse, promote disinformation, heighten polarization, and chill speech.

4.2.2 Justice and retribution

In Kant's (1911/1781) original conception of justice, the need for an institution to administer justice arises from the clear and immediate need to inflict proportionate suffering on an offending individual. This philosophy is known as *retributive justice,* or the belief that offenders deserve sanctions that are proportional to the severity of their crimes. Retributivism is primarily preoccupied with delivering a 'just desert' for a morally wrong act (Kant, 1911/1781), sometimes referred to as "an eye for an eye" (Carlsmith & Darley, 2008; Walen, 2015). Retributive justice, unlike utilitarianism, highlights the need for proportionality in criminal sentencing (Wenzel et al., 2007). For example, in a retributive framework, the death penalty is considered a proportional punishment only for an offender who commits murder.

Retributivist intuitions of moral judgment interact with other theories of justice in complex ways. Carlsmith et al. (2002) argue that even when individuals profess beliefs in the utilitarian-deterrence theory of justice (the belief that a punishment is just only if it effectively discourages others from committing the same crime), they nonetheless continue to apply retributivist assessments to punishment, judging offenders based on degree of moral wrongdoing (Carlsmith et al., 2002). Moral judgment plays a powerful role in retribution and shapes cultural attitudes, policy, and law around appropriate punishments (Giner-Sorolla et al., 2012; Prinz, 2008).

Retributive justice exists within particular social and institutional boundaries, and thus the parameters for what merits retributive punishment are socially constructed and contextual. Indeed, most formal justice systems consider intent when determining punishments. However, different cultures around the world—and even different states in the U.S.—have widely varied beliefs about the appropriateness of some punishments (e.g., death) for criminal offenses. On social media sites, users may seek retribution but have little guidance as to how to enact punishments, or even what an appropriate punishment may be. A widely-known example is that of Justine Sacco, who posted a racist tweet to her 170 followers while boarding a plane to South Africa (Ronson, 2015). Her tweet was captured by mainstream media and resulted in threats of physical and sexual violence and (successful) demands that she be fired. By the time Sacco's flight had landed, the hashtag #HasJustineLandedYet was trending globally on Twitter.

This research seeks to better understand and intervene in online harassment by bridging theories of justice and the underlying circumstances that motivate users to participate in harassing behaviors online. To test the effect of an offense on people's perception of online harassment, we first hypothesize that:

> *H1: Exposure to a retributive prime increases belief that online harassment is a) justified; b) deserved; and c) appropriate.*

Based on the principle of proportionality, or "an eye for an eye," we also hypothesize that participants will view online harassment as more justified and more deserved when the target's perceived offense is demonstrably greater. Second, we hypothesize that:

> *H2: Exposure to a larger retributive prime further increases belief that online harassment is a) justified; b) deserved; and c) appropriate.*

In the Western world, punishment is enacted by a state or institution, and is typically designed to be fair and transparent in process. However, when people take justice into their own hands, it may reflect more individualistic traits and beliefs. Individual people have varied orientations toward retribution; thus, we hypothesize that:

> *H3: Propensity for retributive justice increases belief that online harassment is a) justified; b) deserved; and c) appropriate.*

4.2.3 Social norms and conformity

Social norms—such as values, customs, stereotypes, and conventions—are "social frames of reference" that individuals first encounter through their interactions with others, and which later become internalized (Sherif, 1936). Little is known about how and why norms emerge; however, the widely accepted instrumental theory posits that "norms tend to emerge to satisfy demands to mitigate negative externalities or to promote positive ones" (Hechter & Opp, 2001). Thus, norms are most likely to emerge when they favorably impact a given community's goals (Opp, 2001).

The perceived violation of a social norm is referred to as social deviance. Communities use deviance to establish boundaries—or rather, those who misbehave in turn establish community norms and how rules are made, enforced, and broken (Erikson, 1966; Goode & Ben-Yehuda, 2009). Communities develop norms for appropriateness and enforce those norms through sanctions, both formal (e.g., rules and laws) and informal (e.g., shame or ridicule).

Empirical evidence continues to suggest that group behavior influences individuals to behave similarly. Cialdini (2007) argues that descriptive norms offer "an information-processing advantage," in that by understanding how most people behave in a given situation, a social actor can more quickly decide how to behave themselves (Cialdini et al., 1990). Milgram, Bickman, and Berkowitz's 1969 experiment on the power of crowds is a classic example: when four people standing on a street corner look up at the sky, 80% of passersby will do the same. Normative appeals are most effective when individuals feel connected with a community or group—when we are uncertain about how to behave, we are more likely to "follow the herd," or conform to the perceived norms of a given social group (Goldstein et al., 2008).

Conformity is a type of social influence in which changes in behavior or beliefs are motivated by a desire to adhere to the perceived social norms of a given group. A number of factors increase social conformity, including group size, group cohesiveness, status, self-esteem, and culture. This propensity toward social conformity facilitates distortions of perception (e.g., seeing objects or situations differently than they really are) and distortions of judgment (e.g., believing an act is okay only because other people appear to share that belief). In online environments, factors like relative anonymity, social distance, and social proof may also enhance disposition toward social conformity (Bogardus, 1933; Cialdini, 2001; Walther, 1996). When people witness others engaging in a given behavior, they may seek to conform with the social norms of the group and engage in that behavior themselves. In the context of online harassment, the escalation of threats against a specific individual—sometimes referred to as 'dogpiling'—may be partially explained by the tendency to conform. We hypothesize that:

> *H4: Exposure to conformity increases belief that retributive online harassment is a) justified; b) deserved; and c) appropriate.*

4.2.4 Bystander intervention

Bystander intervention is one potential antidote to undesirable social conformity. The concept of a bystander refers to a person who observes a situation and their subsequent decisions about whether or not to respond or intervene (Darley & Latané, 1968). Intervening in an emergency situation can overcome what is called the bystander effect, where large groups of people observe but ignore offensive behaviors. There are several factors which contribute to the bystander effect, including ambiguity (particularly as emergency situations unfold) and diffusion of responsibility, or an individual's assumption that others are responsible for taking action (or have already done so). Empirical research confirms that the presence of bystanders in an emergency situation reduces helping responses (Fischer et al., 2011).

Promisingly, existing scholarship has also identified several factors that can reduce this bystander effect, including the perceived danger of an emergency, the bystander's relationship to the victim, and the potential risks associated with intervening (Fischer et al., 2011). While group behaviors promote conformity online, propensity toward conformity may be reduced when boundaries around appropriate behavior are questioned. We hypothesize that:

> *H5: Among conforming responses, exposure to bystander intervention decreases belief that retributive online harassment is a) justified; b) deserved; and c) appropriate.*

4.3 Methods

We designed two experiments to test our hypotheses. Both studies were approved by an Institutional Review Board.

4.3.1 Recruitment

Participants for Study 1 were recruited through Twitter. For Study 2, participants were recruited via Twitter and Amazon Mechanical Turk (MTurk). During pilot testing, the survey took an average of 8 minutes to complete; thus, all study participants received $2 as compensation for their time, commensurate with a $15 hourly minimum wage.

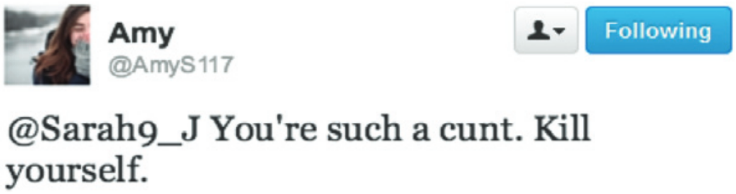
4.3.2 Punishment Orientation Questionnaire

The Punishment Orientation Questionnaire (POQ; Yamamoto, 2014) is an 18-item scale developed to measure individual differences in punishment orientation. In both studies, a participant's score on the POQ's Harsh Retributive Scale (HRS) was used to operationally define their propensity for retributive justice.

4.3.3 Experiment 1: H1, H2, H3

The first study was a 3x1 between-subjects experiment with 3 parts and a total of 35 questions. The first part included a hypothetical scenario of harassment on Twitter and five questions to gauge participants' responses. We chose to simulate a tweet because of the ability for Twitter users to contact people outside of their immediate networks (unlike Facebook, for example, where most interactions occur between Facebook Friends), which would enable someone to engage in retributive harassment regardless of their relationship to the target. The second part of the survey contained the POQ (Yamamoto, 2014), to assess participant's propensity for retributive justice. The final portion of the survey comprised twelve demographic questions (age, gender, race/ethnicity, etc.).

Participants were randomly assigned to one of three conditions: control, low-retributive prime, or high-retributive prime. Participants in the low-retributive prime condition were shown the following prime: “Sarah stole $100 from an elderly couple.” Participants in the high-retributive prime were shown the same information, but with a higher theft amount: “Sarah stole $10,000 from an elderly couple.” Participants in the control condition did not receive a prime. We chose not to include a prime in the control condition—instead of showing a “neutral” prime—because we did not believe a neutral interaction between Sarah (the harassment target) and an elderly couple was possible. In all conditions, participants were shown a harassing tweet sent by Amy to Sarah (see Figure 4.1). Names and avatars were meant to represent white women to control for any possible effects of race and gender.



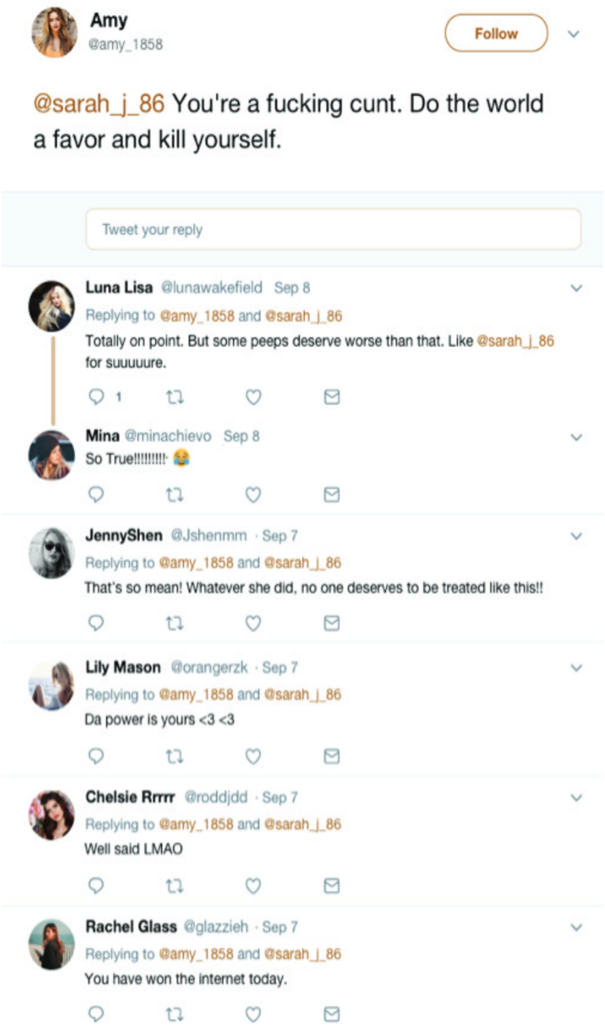
@Sarah9_J You're such a cunt. Kill yourself.

Figure 4.1: Simulated hostile tweet shown to participants.

Participants were asked to rate how *appropriate, deserved,* and *justified* Amy’s tweet to Sarah was on a seven-item Likert scale from absolutely appropriate/deserved/justified to absolutely inappropriate/not deserved/not justified. Participants also responded to two open-ended questions: “If you saw this online, how would you feel?” and “If you saw this online, what (if anything) would you do?”

4.3.4 Experiment 2: H3, H4, H5

The second study also used a 3x1 between-subject design. Participants were randomly assigned into one of three conditions: control; conformity; and conformity + bystander intervention. Participants in all conditions were shown the following information: “Sarah stole $1,000 from an elderly couple.” The survey used the same seven-item Likert scales for appropriateness, deservedness, and justifiability used in study one, with three additional and original measures to understand how the participant would react in certain scenarios: a) “How likely would you be to call out Amy's behavior?" (seven-item Likert scale from extremely unlikely to extremely likely); b) "How likely would you be to call out Sarah's behavior?" (seven-item Likert scale from extremely unlikely to extremely likely); and c) "Whose behavior is more inappropriate?" (a seven-point sliding scale, with Sarah equal to 0 and Amy equal to 7).



Amy @amy_1858

@sarah_j_86 You're a fucking cunt. Do the world a favor and kill yourself.

Luna Lisa @lunawakefield Sep 8
Replying to @amy_1858 and @sarah_j_86
Totally on point. But some peeps deserve worse than that. Like @sarah_j_86 for suuuuure.
1

Mina @minachievo Sep 8
So True!!!!!!!!! 😂

JennyShen @Jshenmm · Sep 7
Replying to @amy_1858 and @sarah_j_86
That's so mean! Whatever she did, no one deserves to be treated like this!!

Lily Mason @orangerzk · Sep 7
Replying to @amy_1858 and @sarah_j_86
Da power is yours <3 <3

Chelsie Rrrrr @roddjdd · Sep 7
Replying to @amy_1858 and @sarah_j_86
Well said LMAO

Rachel Glass @glazzieh · Sep 7
Replying to @amy_1858 and @sarah_j_86
You have won the internet today.

Figure 4.2: Simulated tweets shown to participants in the conformity + bystander intervention condition.

In each condition, participants were presented with a harassing tweet similar to the first study but with some adjusted content. We chose to change the "Following" text to Twitter's "Follow" button, to reduce ambiguity surfaced in the first study's open responses about whether participants knew Amy and Sarah. In the control condition, participants were only shown Amy's harassing tweet. In the conformity condition, participants were shown Amy's harassing tweet with conforming responses (i.e., responses supporting Amy’s harassment of Sarah) from five other users. In the conformity + bystander intervention condition, participants saw Amy’s harassing tweet with conforming responses from five other users, plus one user disagreeing with Amy’s behavior (bystander intervention). We chose to add a sixth reply (see Figure 4.2) to avoid arbitrarily replacing one of the five replies used in the conformity condition. The conformity condition was otherwise identical to the conformity + bystander condition. As in study one, all display names and avatars were meant to represent white women, to control for any possible effects of race and gender.

4.3.5 Experiments 1 and 2: Open Responses

Both surveys included two open-ended questions: “If you saw this online, how would you feel?” and “If you saw this online, what (if anything) would you do?” We used an inductive approach to develop codes (Thomas, 2006). The first author individually read through responses and noted codes by hand. After discussing these initial codes as a research team, we refined a list of codes (35 codes in total).

Resulting codes were organized around several themes, including but not limited to expressions of anger or disapproval toward Amy or Sarah; expressions of sympathy or understanding toward Amy or Sarah; feeling personally upset, offended, amused, or pleased; proportionality (overly harsh or insufficient punishment); expressing a desire to talk to Amy or Sarah, both privately and publicly; and specific actions participants would take if they saw this interaction in their feeds. Two researchers each coded several open responses to test and refine the codebook. In the first study, each open response was independently coded by two members of the research team. Because agreement was high, only the first author coded open responses from the second study. Quotations have been lightly edited for readability.

4.3.6 Participants

For study one, we received 541 total responses from Twitter. For study two, we received 597 total responses (150 responses from MTurk; 447 from Twitter). We removed invalid data from both studies using the following thresholds: a) incomplete responses (i.e., participants who did not reach the end of the survey); b) responses completed in under 200 seconds, which our pilot tests showed to be implausible; c) responses from duplicate IP addresses (all entries were removed); d) responses that had clearly identifiable spam (e.g., entering the word “good” for all open-response questions). For study one, a total of 160 valid cases remained after data cleaning (control group, n=56; low-retributive prime, n=49; high-retributive prime, n=55). For study two, a total of 432 valid cases (143 responses from MTurk; 289 from Twitter) remained after data cleaning (control group, n=145; conformity, n=146; conformity + bystander intervention, n=141).

4.3.7 Data analysis

We used SPSS and R for data cleaning and analysis, using a p-value of .05 for all statistical tests.

Study one: The dataset demonstrated a positively-skewed Poisson distribution, with a majority of the responses falling into either "absolutely inappropriate/not deserved/not justified" or "inappropriate/not deserved/not justified." Between-group one-way Welch's ANOVA was used to compare group mean between the three conditions to adjust for the violation of homogeneity of variance assumption of the standard ANOVA test (Levene's test p<.0001). Similarly, we used a Games-Howell test for post-hoc multiple comparisons due to its robustness against violation of homogeneity of variance. Poisson regression was used to test the relationship between respondents' propensity for retributive justice and their responses (H3).

Study two: This dataset also demonstrated a positively-skewed Poisson distribution. Between-group one-way ANOVA and Tukey's HSD were used to compare means for deservedness. Between-group Welch's ANOVA and the Games-Howell post-hoc test were used for justifiability to adjust for the violation of the homogeneity of variance assumption of the standard ANOVA test (Levene's test p<.0001). Poisson regression was again used to test H3.

4.4 Results

Throughout, we use "offense" to describe the original offense committed by the harassment's target (Sarah's theft). We use "harassing tweet" to describe the retributive harassment targeting the offender (Amy's tweet).

4.4.1 Exposure and magnitude of retributive prime (H1, H2)

The first two hypotheses examined how participants' responses vary when presented with a retributive prime—in which the harassment's target (Sarah) has committed a prior offense (theft)—and whether this priming effect scales with the severity of the offense. H1 states that exposure to a retributive prime would increase the participant's belief that online harassment is justified.

Online harassment of an offender is justified and deserved, but not appropriate. In study one, a between-group one-way Welch's ANOVA revealed a significant difference across priming conditions for *deservedness* (F(2, 92.887)=27.869, p<.001) and *justifiability* (F(2, 93.821)=15.115, p<.001). No significant difference across priming conditions was found for *appropriateness* (F(2, 103.942)=1.620, p=.203). Further, Games-Howell post-hoc multiple comparison suggested that exposure to the retributive prime increased the participant's belief that harassment was *deserved* (M_{High} - $M_{Control}$ =1.864, SEM=.284, p<.001, d=1.254; M_{Low} - $M_{Control}$ =1.564, SEM=.313, p<.001, d=1.013) and justified (M_{High} - $M_{Control}$ =1.246, SEM=.261, p<.001, d=.911; M_{Low} - $M_{Control}$ =1.115, SEM=.292, p<.001, d=.775). In other words, H1 was partially supported: exposure to a retributive prime increases belief that online harassment is justified and deserved. That no significant difference was found for appropriateness suggests that even when Amy's harassment of Sarah was perceived as justified and deserved, participants still recognized that online harassment is not appropriate behavior.

H2 states that the belief of the justifiability of online harassment toward the offender should increase with the severity of the offense, consistent with the retributive value of proportionality. This hypothesis was not supported: no significant difference was found between the \$100 and \$10,000 primes. We specifically used theft as the offense because monetary amounts can be manipulated to be objectively higher or lower; however, it is possible that theft is perceived as a consistently offensive crime, regardless of the amount stolen. Future research should further test this hypothesis with different types of offenses, such as other types of crimes (e.g., vandalism or animal abuse) or social injustice (e.g., racism, white supremacy, or sexism).

4.4.2 Propensity for retributive justice (H3)

H3 states that propensity for retributive justice increases the belief that retributive harassment is justified, deserved, and appropriate. In both studies, a participant's score on the POQ's Harsh Retributive scale (HRS) was used to operationally define their propensity for retributive justice. We used Poisson regression to predict a participant's response to retributive harassment (Amy's tweet) based on their propensity for retributive justice and the priming condition (study one: control, low-retributive prime, and high-retributive prime; study two: control, conformity, and conformity + bystander intervention).

People who favor retributive justice find online harassment of an offender more deserved and more justified. Study one: A likelihood ratio test determined that the proposed model is significant for both *deservedness* (χ2=50.303, p<.001) and *justifiability* (χ2=30.353, p<.001), but not for *appropriateness*. For each one-point increase in the Harsh Retributive scale, there was a 2.9% increase in the participant's response to the *deservedness* of the harassing tweet (B=.029, Deviance=141.049, df=156, Wald χ2=4.751, exp(B)=1.029, p=.029) and a 3.6% increase in the participant's response to the *justifiability* of the harassing tweet (B=.036, Deviance=127.969, df=156, Wald χ2=6.324, exp(B)=1.037, p=.012). In other words, people who have a preference for retributive justice—commonly referred to an "eye for an eye"—believe that online harassment of an offender is more deserved and more justified (but not more appropriate) than do other people.

Study two: As expected (i.e., consistent with results from study one), the proposed model is significant for both *deservedness* (χ2=27.743, p<.001) and *justifiability* (χ2=34.455, p<.001). For each one-point increase in the Harsh Retributive scale, there was a 3.5% increase in the participant's response to the deservedness of the harassing tweet (B=.034, Deviance=529.535, df=428, Wald χ2=17.261, exp(B)=1.035, p<.001) and a 4.6% increase in the participant's response to the justifiability of the harassing tweet (B=.045, Deviance=482.377, df=428, Wald χ2=24.820, exp(B)=1.046, p<.01). We did observe small but significant differences in the mean scores between MTurk (n=143) and Twitter (n=289) responses for *appropriateness* (Twitter M=1.64, SD=1.08; MTurk M=2.42, SD=1.74), *deservedness* (Twitter M=3.15, SD=1.94; MTurk M=4.02, SD=2.20), and *justifiability* (Twitter M=2.54, SD=1.67; MTurk M=3.53, SD=2.13). In other words, MTurk respondents perceived the harassing tweet as more appropriate, more deserved, and more justified than did Twitter respondents. This difference can be partially explained by MTurk respondents' higher scores on the Harsh Retributive Scale (Twitter M=14.24, SD=2.60; MTurk M=17.45, SD=3.07). Future research should assess a wider variety of participants to examine potentially meaningful differences in how users evaluate retributive harassment.

People who favor retributive justice are more likely to call out offensive behavior. In study two, participants were also asked how likely they would be to call out Amy's and Sarah's behavior. These were positively related (r=.51, p<.001), suggesting that people who reported being likely to call out Amy's behavior are also likely to call out Sarah's behavior. Further, participants' propensity for retributive justice was a significant predictor for both: for each one-point increase in the Harsh Retributive scale, there was a 2.3% increase in a participant's reported likelihood to call out Amy's behavior (retributive harassment). Similarly, for each one-point increase in the Harsh Retributive scale, there was a 2.5% increase in participant's reported likelihood to call out Sarah's behavior (theft). This indicates that people who favor retributive justice are more likely to voice public disapproval of offensive behavior.

4.4.3 Conformity and bystander intervention (H4, H5)

H4 and H5 examined the effects of social influence on a participant's perception of online harassment. H4 states that exposure to responses supporting the harassing tweet (i.e., conformity) increases the belief that retributive harassment is justified, deserved, and appropriate. No significant difference across priming conditions was found to support H4. This suggests that individuals' assessments of 'just deserts' may not be easily influenced by others.

Bystander interventions may help prevent dogpiling. H5—that exposure to bystander intervention among otherwise conforming responses should decrease belief that retributive harassment is justified, deserved, and appropriate—was partially supported. No significant effects were found for *appropriateness*. A between-group one-way ANOVA revealed a significant difference across priming conditions for *deservedness* (F(2, 429)=4.247, p<.05). Tukey's post-hoc comparison suggested that, compared to the control group, exposure to bystander intervention among conforming responses decreased the participant's belief that harassment was *deserved* ($M_{\text{conformity+bystander}}$ - M_{Control} =-.688, SEM=.243, p<.05, d=0.329). A between-group one-way Welch's ANOVA also revealed a significant difference across priming conditions for *justifiability* (F(2, 285.405)=4.220, p<.05). Further Games-Howell post-hoc comparison suggested that compared to the control group, exposure to bystander intervention among other conforming responses decreased the participant's belief that harassment was *justified* ($M_{\text{conformity+bystander}}$ - M_{Control} =-.524, SEM=.214, p<.05, d=0.295). In other words, bystander intervention reduces the perception of retributive harassment as justified or deserved.

In study two, participants were asked how likely they would be to call out Amy's and Sarah's behavior. In both cases, neither social conformity nor bystander intervention had a significant effect. Participants were also asked whose behavior was more inappropriate, using a seven-point sliding scale with Sarah equal to 0 and Amy equal to 7 (M=3, IQR=4). For each point increase in Amy's perceived inappropriateness, there was a 6.6% increase in participants' reported likelihood to call out Amy's behavior. However, we did not observe a corresponding effect on participants' likelihood to call out Sarah's behavior—suggesting that across all participants, retributive harassment merits public disapproval in a way theft does not.

4.4.4 Open responses

In both studies, participants were asked to respond to two open-ended questions: "If you saw this online, how would you feel?" and "If you saw this online, what (if anything) would you do?" Open responses are consistent with experimental results but add additional context for interpretation.

Context matters when determining just deserts. In study one's control condition (Amy's harassing tweet with no priming information about Sarah's offense), participants largely expressed that they would be personally upset or offended if they were to see this tweet online. Many participants identified Amy's behavior as being online harassment, which one respondent categorized as "not at all acceptable." However, even in the control condition, some participants expressed a desire to know more context, suggesting that there may be some situations in which Amy's behavior would be justified. Said one participant:

> "Telling someone to kill themselves is inappropriate in any circumstance. The rest [i.e., 'You're such a cunt'] depends on context, which is not available."

Some people said they would try to find out more about Sarah's offense from news websites, or would "read the comments to see how other people feel." Others said they would read through Sarah's or Amy's previous tweets to better assess their overall character.

Harassment can be a proportional punishment, but language matters. Across the low-retributive prime (Sarah stole $100 from an elderly couple) and high-retributive prime (Sarah stole $10,000 from an elderly couple) conditions in study one, participants assessed the proportionality of Amy's sanction based on Sarah's offense and Amy's choice of language. One respondent said that Sarah should have to face consequences for what she did, but that Amy went too far:

> "True, what Sarah did is terrible, and Sarah should have to face consequences for her act. But not death. The harsh judgement reminds me of the KKK and white supremacists, who believe their way is the only way."

Others agreed but maintained that Sarah should be called out for her behavior: "I feel that while confronting Sarah about stealing is the right thing to do, Amy shouldn't have insulted her in that way." Some participants, however, emphasized that "two wrongs don't make a right," and felt time would be better spent assisting the elderly victims of Sarah's crime. Other participants said they would contact the police or seek justice through other means. Said another respondent: "Sounds like Sarah's an asshole, but yelling garbage into the void does nothing to help the wronged."

Still, other participants—particularly in the high-retributive prime condition and in study two—were conflicted. One participant said "given that we know that Sarah stole," they "would not go to bat for her over a potentially over-the-line internet comment." Many participants said while they do not personally condone the language Amy used, they agree with what she said. Said one participant:

> "I think it's wrong to tell people to kill themselves (obviously), but who steals $10,000 from an elderly couple? She should be told in no uncertain terms what a horrible person she is."

Some participants were not at all conflicted by Amy's language, and said that if they were to see this tweet online, they would laugh, like or retweet Amy's tweet, or be otherwise amused. Others said they would "join in the bashing." One participant in study two applauded Amy for calling out Sarah's behavior: "At least some kids still have morals these days, even if they have foul mouths."

Online harassment has become normalized—and intervention is risky. Although some participants said they would report Amy's tweet for harassment or would message Sarah to offer emotional support, most participants said they would do nothing. One participant, when asked what (if anything) they would do, said: "Ignore it. It's not my battle." Other participants said they would react differently depending on whether or not they personally knew either of the women. Several participants who did not agree with Amy's behavior said they would not call her out or otherwise intervene, for fear of facing harassment themselves. Said one participant in study two: “In this day and age, I would be afraid to intervene.” Another said: “Honestly, I probably wouldn’t do anything—any direct response just opens you up to that kind of vitriol.”

Across both studies, participants felt that online harassment was becoming normalized. One respondent said they don’t feel it’s ever appropriate to tell someone to kill themselves, but they felt desensitized to seeing these types of sentiments expressed online:

> “I honestly feel so desensitized to responses like Amy's. They are everywhere. I wouldn't feel much of anything—other than rolling my eyes and moving on.”

Another respondent agreed: “It doesn't look like Amy is serious, so I'd shrug it off as typical Twitter hyperbole over Sarah's admittedly atrocious behavior.” Participants directly associated this feeling of normalization with their unwillingness to report, call out, or otherwise intervene in Amy’s harassment of Sarah. Said one participant: “I wouldn't do anything in particular. The internet is an awful place.”

4.5 Discussion and Future Work

4.5.1 Designing technologies to encourage bystander action

Our results show that online harassment is perceived to be more justified and more deserved, but not more appropriate, when the target has committed some offense. Promisingly, exposure to a bystander intervention among other conforming responses decreased this perception—suggesting that designs encouraging bystander action could discourage harassment through normative enforcement.

Platforms can encourage bystanders to intervene by reducing ambiguity and diffusion of responsibility, factors which contribute to bystander apathy (Darley & Latané, 1968). Indeed, recent research suggests that bystanders are motivated to intervene when they understand the breadth and impact of harassment, factors which are obscured in distributed, cue-sparse environments (Blackwell et al., 2017). Experimental research confirms that bystanders feel more personally responsible, and are more likely to intervene directly, when exposed to multiple instances of harassment targeting a single user (Kazerooni et al., 2018). Given these findings, social media platforms should counteract ambiguity by making the harmful impacts of online harassment more visible. Further, although many current interventions aim to obscure or hide harassment from both targets and bystanders (e.g., blocklists; block and mute tools; Twitter's "Quality Filter"), reminding potential bystanders that online abuse is both prevalent and inappropriate could foster a greater sense of personal responsibility.

Online bystanders are more likely to intervene in indirect ways (e.g., by reporting content to platforms) than by responding directly to perpetrators, due to the social and physical risks of direct intervention (Dillon & Bushman, 2015). Platforms should prioritize simple, indirect interventions that do not put bystanders at risk. For example, HeartMob (iheartmob.org) provides bystanders with specific and private ways to take action, such as sending a supportive message or documenting abuse (Blackwell et al., 2017). Finally, some participants said they would look to other users' responses to determine how they themselves should act (i.e., descriptive norms); anonymously highlighting bystander interventions when they do occur may encourage other users to do the same.

4.5.2 Designing technologies to mitigate retribution

Our qualitative data suggests that some people censor themselves due to fear of retribution, suggesting that retributive harassment may contribute to chilled speech online. This could be particularly damaging for marginalized populations, including women, people of color, and LGBT people, who are already more likely to censor themselves online because they fear facing harassment (Duggan et al., 2014; Lenhart et al., 2016; Rainie et al., 2017). A danger of retributive harassment, and its widespread use, is that marginalized voices will be silenced while socially dominant perspectives are amplified.

Because the affordances of existing social media platforms exacerbate retributive harassment—and also limit potential consequences for those who choose to engage in vigilantism—we should instead consider designing platforms that encourage alternative forms of justice-seeking. An emerging alternative to retributive justice is *restorative justice*, which prioritizes improving society for the future. Restorative justice provides a voice to both victim and offender: the victim is encouraged to express a willingness to forgive, and the offender is encouraged to accept responsibility for their actions, with the goal of mending conflicts between individuals and communities (Wenzel et al., 2008).

Future work should explore ways of integrating restorative approaches into the design of online communities. Social media platforms could algorithmically detect surges of retributive harassment and experiment with designs that introduce mediation, reconciliation, and proportionality. This might involve the use of deescalating language that draws on shared experiences and understanding, mechanisms that enable or require social resolution, or the creation of spaces where communities can voice their feelings and concerns (Simonson & Staw, 1992). For example, a new type of temporary Facebook Group could serve as a moderated platform for communities to work together with offenders to re-establish and validate relevant community values, restoring justice through social consensus (Wenzel et al., 2008). If social media platforms were to leverage their existing community features to encourage restorative mediation, justice could be restored without the use of retributive sanctions—promoting civil and inclusive participation online by enabling reconciliation at scale.

4.6 Conclusion

We propose the concept of *retributive harassment* to explain why and how certain kinds of online harassment occur—namely, when online harassment is used as a controversial form of social sanctioning. We reflect on the affordances of social media platforms that enable retributive harassment, and we advocate for the design of systems that encourage more restorative forms of justice-seeking.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant #1318143. We thank Cass Wilkinson Saldaña for contributions to the research design.

Chapter 5 Content Moderation Futures

5.1 Abstract

This study examines the failures and possibilities of contemporary social media governance through the lived experiences of various content moderation professionals. Drawing on participatory design workshops with 33 practitioners in both the technology industry and broader civil society, this research identifies significant structural misalignments between corporate incentives and public interests. While experts agree that successful content moderation is principled, consistent, contextual, proactive, transparent, and accountable, current technology companies fail to achieve these goals, due in part to exploitative labor practices, chronic underinvestment in user safety, and pressures of global scale. I argue that successful governance is undermined by the pursuit of technological novelty and rapid growth, resulting in platforms that necessarily prioritize innovation and expansion over public trust and safety. To counter this dynamic, I revisit the computational history of care work, to motivate present-day solidarity amongst platform governance workers and inspire systemic change.

5.2 Introduction

Online platforms are essential venues for social interaction. As social media platforms[1] grow in size, reach, and influence, the question of how to govern human behavior at scale becomes increasingly critical. Contemporary platform governance hinges on what Roberts (2016) deems *commercial content moderation*, a complex assemblage of human labor, automated technologies, and organizational practices intended to ensure user safety through the scaled enforcement of standardized rules. Despite their growing prominence, scaled content moderation systems are frequently characterized by opacity, inconsistency, and ineffectiveness,

[1] Tarleton Gillespie (2018) defines social media platforms as "online sites and services that (a) host, organize, and circulate users' shared content or social interactions for them, (b) without having produced or commissioned the bulk of the content, (c) built on an infrastructure beneath that circulation of information, for processing data for customer service, advertising, and profit."

particularly for users whose identities, histories, and experiences are not represented by dominant ideological structures.

The rise of social media platforms has led to unprecedented challenges in moderating human behavior at scale, with billions of users generating posts, images, and videos daily. Platforms like Facebook, YouTube, TikTok, and Reddit face mounting pressure from governments, advertisers, and users to prevent harm while also protecting expression. Though numerous scholars have examined the sociopolitical implications of content moderation—from the algorithmic biases embedded in automated enforcement tools (Noble, 2018) to the material conditions of outsourced moderators (Roberts, 2016; Roberts, 2019)—comparatively little research has engaged directly with the broad spectrum of professionals responsible for implementing and interrogating scaled platform governance. While content moderation professionals—including both industry and civil society workers, spanning roles across policy, operations, research, engineering, and other relevant domains—possess deep, practical knowledge of the governance systems they help enact, their experiences remain largely absent from both public and scholarly discourse.

Through a series of participatory design workshops with 33 content moderation professionals, this research surfaces a situated understanding of the values, challenges, and contradictions that shape contemporary social media governance. While experts agree that successful content moderation is principled, consistent, contextual, proactive, transparent, and accountable, technology companies repeatedly fail to achieve these goals at scale. The impractical pursuit of universal solutions to global governance yields vague, Western-centric policies, marginalizing non-dominant perspectives in ways further compounded by the technical limitations of automated enforcement systems. Business incentives conflict with user safety goals, resulting in chronic underinvestment and exploitative labor practices—and leaving individual workers feeling overburdened and isolated from their closest peers. I argue that successful governance is undermined by the pursuit of technological novelty and rapid growth, resulting in platforms that necessarily prioritize innovation and expansion over public trust and safety. I conclude by revisiting the computational history of care work, positioning worker solidarity as the foundation for meaningful structural change.

5.3 Related work

5.3.1 Failures of contemporary social media governance

Platforms rely on a combination of human labor and automated tools to enforce policies and guidelines, adhere to regulatory compliance obligations, and manage reputational risk (Gillespie, 2018; Roberts, 2019; Tyler et al., 2025). Social media companies typically publish formal policies for appropriate use, such as Meta's Community Standards[2], which describe specific behaviors that are and are not allowed on Facebook, Instagram, Messenger, and Threads. Users whose behavior is determined to violate these policies are subject to a variety of sanctions, including content removal and temporary or permanent account suspension.

Enforcing these policies at a global scale has proven to be an impossible challenge (Gillespie, 2018; Roberts, 2019). Many social media users are not aware that rules exist; a user is often first exposed to platform rules when they have violated one, often inadvertently (Chandrasekharan et al., 2018; Katsaros et al., 2022; Tyler et al., 2025). Other users may have an ambient awareness of the existence of site policies, but few can articulate specific rules or examples of inappropriate behavior, a problem intensified by the lack of transparency social media companies provide into their specific enforcement practices (Gillespie, 2018; Suzor, 2019). Rules also vary widely across platforms and often invoke complex or specialized terminology (Pater et al., 2016; Jiang et al., 2020), further exacerbating the burden on individual users to sufficiently understand platform rules and any consequences of breaking them.

These private rules—which regulate users' speech, conduct, and access—are enforced through scaled content moderation practices, or the systems and processes used to monitor and evaluate large volumes of user-generated text, images, and videos (Gillespie, 2018; Roberts, 2019). Contemporary platform governance ecosystems rely heavily on *commercial content moderation* (Roberts, 2016), or the process of outsourcing scaled content moderation labor to external service providers in lower-wage regions of the Majority World. A typical commercial content moderator reviews hundreds of pieces of content per day, earning low wages despite unrealistic performance standards and constant exposure to objectionable content (Roberts, 2016; Roberts, 2019). Moderators are expected to make individual enforcement decisions in a matter of seconds, requiring memorization of complex policies that frequently change.

[2] http://transparency.meta.com/policies/community-standards/

While users can report content directly to platforms for potential review (Crawford & Gillespie, 2016; Tyler et al., 2025), many technology companies leverage machine learning algorithms to automatically detect—and in some cases, enforce against—potentially violative content (West, 2018; Gorwa et al., 2020; Vaccaro et al., 2020). Developing such models requires a similarly invisible workforce to generate sufficiently large volumes of labeled training data (Gray & Suri, 2019), with platforms like MTurk facilitating the on-demand employment of data labeling workers, whose labor is required to translate complex cultural context into a format "legible" to computers (Irani, 2023).

5.3.2 Promoting social justice under platform capitalism

The purported promises of scaled content moderation have been increasingly undermined as social media platforms repeatedly fail to adequately prevent harm, both to individuals and society at large. Interdisciplinary scholarship has highlighted numerous ways in which scaled moderation systems fail, both due to insufficient technical systems and the structural, economic, and epistemological assumptions embedded within them (Gillespie, 2018; West et al., 2019; Gorwa et al., 2020). Despite increasing public pressure to discourage and even limit discriminatory or otherwise threatening speech (Grimmelmann, 2013), many social media platforms continue to operate under a guise of neutrality, refusing to reckon with or even acknowledge their role in actively shaping contemporary culture (Chander & Krishnamurthy, 2018; Gillespie, 2018).

When platform governance fails, social media users are exposed to a variety of sociotechnical harms (Schoenebeck & Blackwell, 2021; Domínguez Hernández et al., 2023). Online harassment, hate speech, and other forms of interpersonal abuse inflict emotional, psychological, and physical distress (Blackwell et al., 2017; Vitak et al., 2017; Im et al., 2022). Rumors, conspiracy theories, and state propaganda distort public understanding and promote radicalization, sometimes culminating in physical violence (Marwick & Lewis, 2017; Starbird et al., 2019; Marwick et al., 2021). Suicide and other forms of self-harm are normalized or even glorified (Chancellor et al., 2016; Pater & Mynatt, 2017).

These harms are compounded by platform designs that perpetuate existing structural inequities, resulting in disproportionate impacts to vulnerable social media users, including women, people of color, queer people, transgender people, refugees, dissidents, and many others (Blackwell et al., 2017; Caplan, 2018; Noble, 2018; Pearce et al., 2018; DeVito et al., 2021).

Platform governance practices further exacerbate these harms, with marginalized users facing disproportionate content moderation experiences (Haimson et al., 2021; Lyu et al., 2024; Mayworm et al., 2024; Thach et al., 2024), due in part to imprecise automated enforcement mechanisms that collapse social and political complexity (West et al., 2019; Gorwa et al., 2020). Despite these disparities, platforms offer users limited opportunities for recourse (West, 2018; Vaccaro et al., 2020).

Technology companies currently wield considerable power over public discourse, with limited accountability for, or transparency into, their specific governance practices (Suzor et al., 2019; Keller & Leerssen, 2020). Though social media platforms have failed to adequately regulate themselves, global regulators have also struggled to effectively respond, in part because platforms, who operate across borders, must navigate a patchwork of local and national laws (Klonick, 2017; Suzor, 2019; Keller & Leerssen, 2020). Suzor (2019) argues that platform governance amounts to private regulation by proxy, often substituting for formal legal mechanisms. As platforms adopt increasingly sophisticated governance regimes, such as Meta's Oversight Board (Klonick, 2019), distinctions between private and public authority blur, further disempowering individual users.

In response to these concerns, novel regulatory frameworks are emerging, though unevenly across jurisdictions. The European Union's Digital Services Act (DSA) and General Data Protection Regulation (GDPR) represent efforts to impose transparency and accountability on social media platforms, particularly regarding content moderation and data usage (Suzor, 2019; Keller, 2022). In contrast, regulatory responses in the United States remain fragmented and largely deferential to corporate self-governance, constrained by First Amendment protections and historical reluctance toward state regulation of speech (Balkin, 2021; Keller, 2023). As social media platforms become increasingly central to contemporary public life, how they are governed—by whom, for whom, and through what mechanisms—remains a critical challenge.

5.3.3 Trust & Safety and the "Techlash"

Social media governance has become increasingly professionalized in recent years, largely under the umbrella of "Trust & Safety" (T&S), a common industry term used to describe various people, policies, and products that broadly support user safety. Though the specific origins of the term are unknown, "trust and safety" (and similar variants) have been used throughout the technology industry since at least 1999, with eBay referenced as one early adopter (Boyd, 2002; Cryst et al., 2021).

Contemporary technology companies have largely adopted what Zuckerman and Rajendra-Nicolucci (2023) describe as a “customer service” model of governance, adapting familiar bureaucratic structures for efficient, centralized management of customer concerns to the emergent challenges associated with unanticipated community growth. The increasing professionalization of online governance also emerged in response to growing legal liability, as modern laws—such as the Digital Millennium Copyright Act (DMCA)[3] of 1998—required technology companies operating in the United States to establish formal procedures for responding to potentially infringing content.

Trust & Safety teams, “most often born in a crisis” (Maxim et al., 2022), are formalized over time as companies mature, alongside the evolution from ad hoc governance decisions to explicit policies and processes (Zuckerman & Rajendra-Nicolucci, 2023). While smaller companies may still operate more “artisanal” approaches to platform governance (Caplan, 2018), content moderation practices have rapidly industrialized, with Meta reporting a combined “safety and security” team of 40,000 people (Meta, 2025b). Globally, the Trust & Safety Professional Association (2025) estimates over 100,000 T&S professionals employed in a variety of functions and roles, including policy, operations, and compliance teams, as well as safety-focused roles in product, research, and engineering departments.

The Trust & Safety Professional Association (TSPA)[4], established in 2020 alongside the Trust & Safety Foundation[5] (Goldman, 2020), is one of several recent institutions designed to support, structure, and standardize the T&S profession, including TrustCon[6] (a professional conference first hosted by TSPA in 2022) and the Journal of Online Trust & Safety[7], an academic journal of peer-reviewed research first introduced in 2021 (Cryst et al., 2021). Still, the structure of Trust & Safety teams varies substantially across companies, in part because the governance of a specific platform is informed by its unique characteristics and challenges—but

[3] Pub. L. No. 105-304, 112 Stat. 2860 (Oct. 28, 1998).

[4] http://www.tspa.org

[5] http://www.trustandsafetyfoundation.org

[6] http://www.trustcon.net

[7] http://tsjournal.org

also because T&S functions are often formed extemporaneously, with even founding members of fledgling Trust & Safety teams sometimes lacking relevant prior experience (Maxim et al., 2022; Tyler et al., 2025). This helter-skelter assemblage of professional practice has profound consequences for platform governance, as the present research will demonstrate.

Recent years have brought significant changes to the burgeoning Trust & Safety industry, most notably following the October 2022 acquisition of Twitter (since rebranded as "X") by Elon Musk, best known for his temporary involvement in the second Trump administration's Department of Government Efficiency (DOGE). During his brief but tumultuous tenure with DOGE, Musk implemented aggressive "efficiency" reforms—including dramatic reductions in staffing and several agency closures—that closely resembled his early management of Twitter, purchased in a $44 billion deal he famously attempted to terminate (Schiffer, 2024). In what some have described as "the most controversial corporate takeover in history" (Mezrich, 2022), Musk immediately cut Twitter's workforce in half—including nearly a third of Twitter's former Trust & Safety team—and reinstated the accounts of more than 6,000 previously banned users, including conspiracy theorist Alex Jones (Basic Online Safety Expectations, 2024; Tyler et al., 2025). In the weeks following Musk's takeover, the average proportion of hate speech on the platform quadrupled (Hickey et al., 2023; Schiffer, 2024).

Many other major technology companies have pursued similar reductions to safety programming in the intervening years, including large-scale workforce reductions—reigniting growing public criticism widely described as the "Techlash" (Su et al., 2021; Helles & Lomborg, 2024). Nearly 760,000 people have been laid off by global technology companies since March 11, 2020—when COVID-19 was first declared a pandemic by the World Health Organization (WHO)—with more than 660,000 of those layoffs occurring since April 14, 2022, the date Musk publicly announced his unsolicited offer to purchase Twitter (Lee, 2020). In January 2025, ahead of the second presidential inauguration of Donald Trump, Meta announced substantial changes to its enforcement policies and products—including ending third-party fact-checking, relocating Trust & Safety teams from California to Texas, and significantly reducing automated content detection and demotion—under the guise of "free expression" (Meta, 2025a). Though company profits continue to soar, technology workers are left demoralized by systemic issues they lack the power to sufficiently address (Su et al., 2021).

5.3.3.1 Situating platform governance in worker experiences

While scholars across disciplines have devoted significant attention toward understanding both informal and formal mechanisms for regulating online behavior (e.g., Lampe & Resnick, 2004; Lessig, 2006; Hardaker, 2010; Diakopoulos & Naaman, 2011; Sternberg, 2012; Kiesler et al., 2012; Cho & Acquisti, 2013; Marwick & Miller, 2014; Guberman, Schmitz, & Hemphill, 2016; Massanari, 2017; Blackwell et al., 2018; Im et al., 2022; Han et al., 2023), fewer studies directly examine the experiences and perspectives of the network of professionals—business leaders, government employees, technology workers, and so on—responsible for enacting platform governance at scale.

Scholars who have directly examined platform governance through a worker-centric perspective (Roberts, 2016; Gray & Suri, 2019; Roberts, 2019; Ruckenstein & Turunen, 2020) primarily investigate the experiences of commercial content moderators, or the shadow workforce of outsourced workers responsible for evaluating large volumes of user-generated content in accordance with platform policies and global laws (Roberts, 2019). While understanding the experiences of platform governance's most disenfranchised workers is certainly critical to identifying and advancing opportunities for meaningful change, situating platform governance systems within a broader range of professional practices may produce a more comprehensive understanding of their ideological and operational underpinnings, a necessary prerequisite for systemic transformation.

Direct inquiry of the workers responsible for enacting these technologies is rare, in part, by design. Companies impose stringent (but questionably enforceable) non-disclosure agreements, while individual employees navigate increasing job security concerns and broader threats to their psychological safety amidst frequent, often sudden layoffs (Lee, 2020). As such, most scholarly investigations of platform governance rely primarily on assumptions about professional practices. In the present research, I seek to structure a shared understanding of platform governance across multiple competing perspectives, through direct inquiry of the experiences, perspectives, and interpretations of several categories of differently positioned workers (Suchman, 1995) generally described as *content moderation professionals*. By making this work more visible, I aim to produce a more intimate understanding of the complex landscape of people, practices, and politics that ultimately determines how platforms are governed—which may conflict with companies' platform governance intentions, regulators' platform governance expectations, and the broader public's platform governance assumptions.

Influenced by a growing body of human-computer interaction (HCI) and computer-supported cooperative work (CSCW) literature promoting worker-centric inquiries of technological progress (Irani & Silberman, 2013; Fox et al., 2020; Su et al., 2021; Wolf et al., 2022), this research aims to broaden current understandings of platform governance by examining the lived experiences of practitioners who enact, examine, and engage with these same sociotechnical systems. The results—which both corroborate and complicate existing assumptions about workers' underlying practices—reflect insights from workers with a broad range of both theoretical and practical platform governance expertise, from part-time content moderators and professional researchers to corporate vice presidents.

5.3.3.2 Situating platform governance in broader social worlds

Still, the experiences of individuals—the traditional focus of qualitative inquiry—are only one component of the complex ecosystem of actors that inform and influence contemporary platform governance. Despite technosolutionist promises of orderly, frictionless futures, our social realities are never simple or straightforward, and investigating the "mess" of human complexity (Dourish & Bell, 2011) is necessary to construct realistic paths toward progress. I leverage situational analysis (Clarke, 2005) to ground my interpretation in the broader "situation" or social ecology of social media governance, which necessarily implicates collective actors from multiple social worlds, often with competing interests. By examining the sociopolitical experiences of content moderation professionals in juxtaposition with the institutional processes and power imbalances that construct their discursive realities (Foucault & Nazzaro, 1972), this study seeks to examine the underlying logics that construct and sustain contemporary platform governance and articulate aspirational, worker-centered visions for more equitable technology futures.

Contemporary social media governance is primarily constructed by four social worlds (see Figure 5.1), representing the interests of businesses, states, the public, and the working class. A social world represents a group with broadly shared commitments and practices, each comprising various organizations of actors who participate in platform governance to varying degrees, and with differential access to social and political power. Business interests are the dominant bloc of platform governance, with technology companies themselves holding the majority of power. This is particularly (but not only) true of technology companies with a public-facing social platform product, such as Meta, TikTok, and X, who maintain control over platform design, policy implementation, user data, and technical infrastructure.

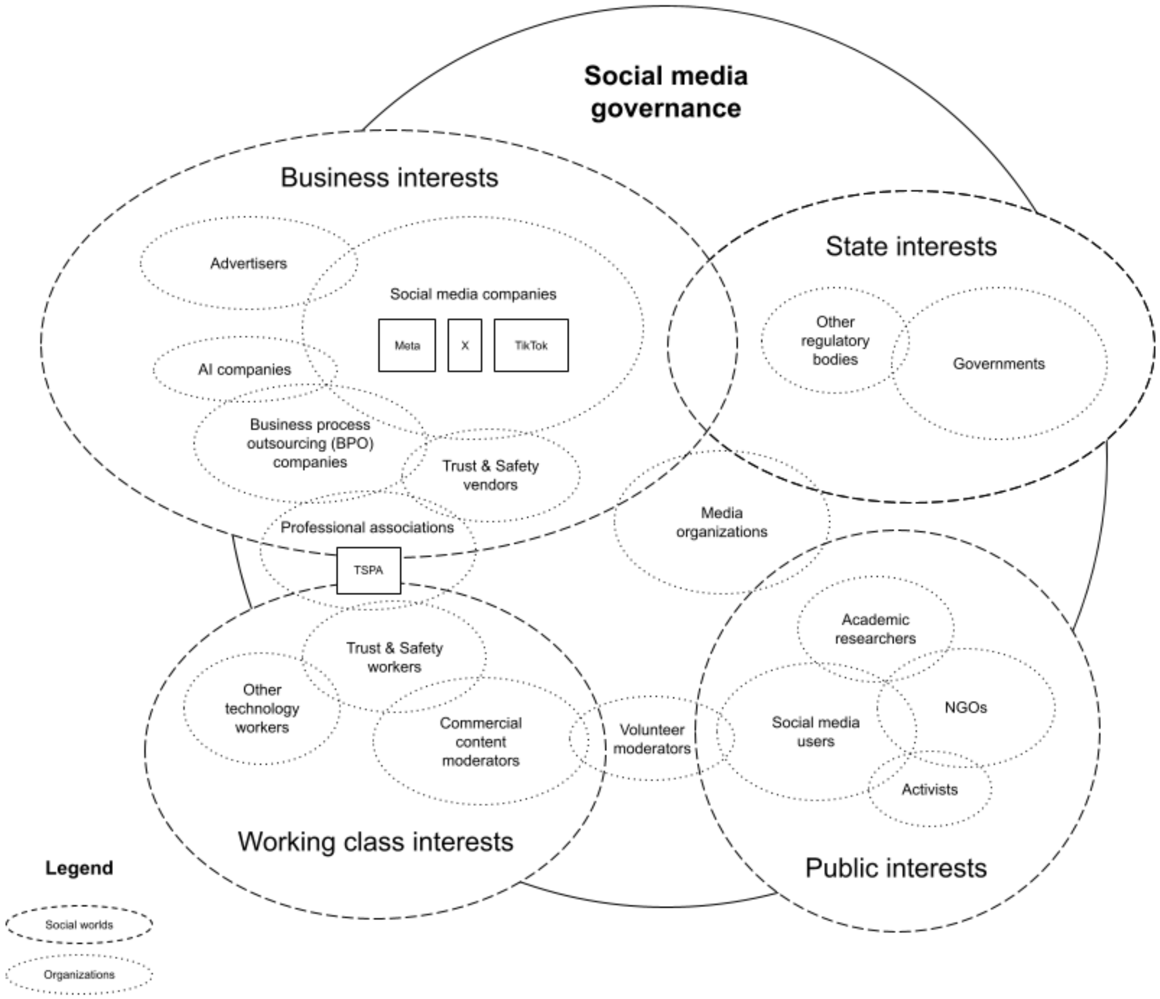


Figure 5.1: A social worlds map illustrating the primary collective actors who construct and maintain contemporary platform governance, including both formal and informal organizations.

Other technology companies also participate in enacting platform governance: business process outsourcing (BPO) companies (such as Accenture, Cognizant, and TaskUs) manage critical content moderation, data labeling, and customer support tasks; artificial intelligence (AI) companies (such as OpenAI, Hive, and Spectrum Labs) develop technologies used to augment or automate content review processes, including machine learning models designed to detect potentially violative or otherwise problematic content; other vendors provide various Trust & Safety services, including content moderation platforms (such as Cinder and Checkstep), identity verification services (such as Persona and Prove), and risk intelligence tools (such as Crisp and CrowdStrike). Also representing business interests are advertisers, who participate in platform governance primarily through their relationships with social media companies, which generate the vast majority of their revenue through targeted advertising.

Businesses are predominantly interested in profitability, in direct conflict with the interests of the working class, whose devalued labor produces profit (Marx, 1844/1959). Working class interests broadly include stable employment, safe working conditions, and fair pay, though class stratification results in privilege and power disparities between workers (Ehrenreich & Ehrenreich, 1977). In the context of social media governance, commercial content moderators—who are typically employed by BPOs in the Majority World, often on a contract basis—are the most structurally disempowered, despite directly managing the daily labor of platform governance. Other technology workers, including those in Trust & Safety and related roles, generally enjoy significantly higher wages and more direct proximity to corporate decision-making. However, as the present research will demonstrate, most technology workers lack the necessary structural power to effectively influence platform governance, confirming the utility of positioning salaried technology employees in broad coalition with their more severely disenfranchised peers. All workers are subject to exploitation and face increasing labor precarity as pervasive automation increases (Riek & Irani, 2025).

State interests, represented by governments and other regulatory bodies, largely influence social media governance through the direct application of formal laws. Though legal requirements vary across jurisdictions, recent regulations include laws governing the visibility of certain content (such as Germany’s Network Enforcement Act, 2017); laws protecting user privacy (such as California’s Consumer Privacy Act, 2018); and laws mandating corporate transparency (such as the European Union’s Digital Services Act, 2022). Certain regulations have outsized influence on contemporary platform governance, such as 47 U.S.C. § 230, a provision of the United States’ Communications Decency Act (1996) under which online service providers enjoy limited liability for user-generated content (Gillespie, 2018; Citron & Franks, 2020). State actors may also exert informal or covert influence over platform governance—for example, by pressuring platforms to remove politically sensitive content (Park & Sang, 2023) or sponsoring influence operations designed to manipulate public opinion (Starbird et al., 2019). As in other social worlds, differences in power and proximity to capital determine overall influence, and wealthier or more geopolitically powerful nations may shape platform governance in ways that further marginalize the interests of the Majority World.

Public organizations also inform social media governance, though public interests are generally more dispersed across loosely organized groups of social actors, including activists, academics, and social media users themselves. No organization is a monolith, and the interests of individual social media users clearly vary; however, public interests can be broadly understood as representing autonomy, privacy, and general welfare. Because they lack the institutional power present in other social worlds, public social actors—including people who use, or are otherwise proximate to, technology products and services—are among the most marginalized by existing platform governance structures. While advocacy by non-governmental organizations (NGOs) and other civil society groups can highlight technology harms and recommend potential solutions (Severance, 2013), public actors typically lack structural mechanisms for motivating or otherwise enforcing social progress. Without thoughtful, transparent, and equitable governance, social media platforms risk capture by the interests of powerful and well-resourced actors, rather than serving the interests of the broader public.

5.4 Methods

Data was collected during six participatory design workshops held via Zoom in June and July 2022. Participants were not compensated for their participation in this study. Potential participants were recruited via judgment sampling, based on two factors:

1. *Expertise.* Expertise was the primary factor in participant recruitment, as the study required participation from experts who specialize in professional content moderation and platform governance. I specifically recruited experts with a wide range of experiences, including professionals employed in the technology industry and in broader civil society.

2. *Relational trust.* A second factor, particularly for the recruitment of potential industry participants, was a desire to initially contact only those experts with whom the author had some pre-existing relationship, for the safety and comfort of participants as well as the author. Participating in a research study as a representative of a company presents significant risk, particularly in the current corporate technology climate. Industry media coverage increasingly highlights privileged information from presumably internal

sources, resulting in fear, distrust, and concern for potential retaliation or job loss. The author did not wish to risk anyone's career, reputation, or material security should someone view participation in this study as a breach of corporate confidentiality.

Recruitment continued until saturation was reached, resulting in a total of 33 participants with a collective 230 years of content moderation experience. Participants' professional content moderation experiences ranged from those working for technology companies (including Airbnb, Discord, Google, Instagram, Lyft, Meta, Microsoft, and Twitter; represented as "I" in Tables 5.1 and 5.2, indicating industry experience) to those working for universities (including Carnegie Mellon University, Cornell University, Georgia Institute of Technology, Northeastern University, Rutgers University, Stanford University, University of California Irvine, University of California Los Angeles, University of California San Diego, University of Colorado Boulder, University of Illinois Urbana-Champaign, University of Maryland, University of Michigan, and Yale University; represented as "A" in Tables 5.1 and 5.2, indicating academic experience) and non-profit institutions (including AI Now Institute, American Library Association, Dangerous Speech Project, Data & Society, and Meedan; represented as "C" in Tables 5.1 and 5.2, indicating other civil society experience). Nine participants had also been content moderators themselves, either professionally or in a volunteer capacity (represented as "M" in Tables 5.1 and 5.2). One participant served as a government advisor (represented as "G" in Tables 5.1 and 5.2).

At the time of recruitment, participants had between 2 and 20 years of experience in content moderation specifically. While the median number of years' experience across all participants is 6, participants who chose to be identified in the resulting manuscript represent more experience on average (median 8 years) than participants who chose to participate anonymously (median 5 years), which suggests that more tenured professionals may be able to discuss their experiences and perspectives more openly than their junior peers. Similarly, most industry participants chose to participate anonymously: of the 16 identified participants, only 7 report industry-specific experience (median 12 years), including 6 former employees of major technology companies (3 of whom were subsequently employed by academic or other non-corporate institutions; one participant was working as an independent consultant, and the remaining two participants were between jobs). Only one identified participant (P8) was employed by a major technology company (Meta) at the time of data collection. In contrast, 11 of the 17 anonymized participants were current full-time employees ("FTEs") of major technology companies (median 5 years). Restrictive non-disclosure agreements and other systemic barriers prevent many workers in the technology industry from openly discussing their experiences and knowledge, even in a research context.

During the workshops themselves, most participants chose to identify themselves to other workshop participants; some chose to participate under a pseudonym. Participants were also afforded the ability to self-identify in the resulting manuscript, at each individual's desired level of identifiability (e.g., with or without a name or institutional affiliation). While participants were invited to participate in the workshops with or without video, all participants were required to join from a laptop or other device that would allow them to view a shared screen. Participants were also given links to workshop documents (e.g., Google Slides), should they wish to contribute directly to the documents themselves. Relevant documents were secured after each workshop concluded. Workshops were recorded (voice, screen, and chat) and transcribed for research purposes only. Participants were able to change their level of identifiability or withdraw from the study at any time before the paper's publication.

Two workshops had seven total participants; two workshops had four participants; one workshop had five participants; and the final workshop had six participants. In order to more appropriately examine the broader social ecology of platform governance, each workshop included content moderation professionals from various social worlds, to surface the perspectives of diverse individuals while also facilitating interactions between them. Workshops were structured as a *future workshop*, a participatory design method used to collaboratively define an ideal future outcome, without existing resource restraints, technical limitations, and organizational realities (Jungk & Müllert, 1987; Vidal, 2006; Hardy et al., 2022). After orienting around a collective vision for a better future, participants determine potential solutions and procedures to advance the status quo closer to this imagined future. The goal of a future workshop is to bring together a diverse group of people who share interest in a common problem—in this case, scaled content moderation. The structure of a future workshop allows participants to identify both shared and conflicting goals, which can then be used to structure a collective path forward. A future workshop has three stages: 1) *Inspiration* (identify common problems; e.g., reflect on the present-day situation), 2) *Ideation* (generate a shared vision for a better future; e.g., what would the ideal solution be, assuming anything is possible?), and 3) *Implementation* (discuss and prioritize potential ideas and solutions).

Table 5.1: Identified participants (self-described), including academic (A), other civil society (C), government (G), technology industry (I), and moderator (M) experiences at time of data collection (2022).

	Name	**Role(s)**	**Experience**
P1	Susan Benesch	Executive Director, Dangerous Speech Project	9 years (C)
P2	Robyn Caplan	Senior Researcher, Data & Society Research Institute *(formerly)* Researcher, Rutgers University	6 years (A,C)
P3	Anne Disabato	*(formerly)* Product Manager, Facebook *(formerly)* Product Manager, Twitter	7 years (I)
P4	Casey Fiesler	Associate Professor, University of Colorado Boulder Moderator	5 years (A,M)
P5	Eric Gilbert	Associate Professor, University of Michigan *(formerly)* Associate Professor, Georgia Institute of Technology	7 years (A)
P6	Sarah Gilbert	Research Manager, Citizens and Technology Lab, Cornell University *(formerly)* Researcher, University of Maryland Moderator	6 years (A)
P7	Amber Grandprey-Shores	Moderator, Twitch Moderator, Discord	5 years (M)
P8	Mark Handel	Researcher, Meta	10 years (I)
P9	Del Harvey	Independent consultant *(formerly)* Vice President, Trust & Safety, Twitter	20 years (I)
P10	Matthew Katsaros	Director, Social Media Governance Initiative, Yale University *(formerly)* Researcher, Twitter *(formerly)* Researcher, Facebook	11 years (A,I)
P11	Kat Lo	Content Moderation Lead, Meedan *(formerly)* Researcher, Instagram *(formerly)* Researcher, University of California Irvine Moderator	14 years (A,C,I,M)
P12	Sarah T. Roberts	Associate Professor, University of California Los Angeles *(formerly)* Staff Researcher, Twitter *(formerly)* Fellow, American Library Association	12+ years (A,C,I)
P13	Joseph Seering	Researcher, Stanford University *(formerly)* Researcher, Carnegie Mellon University	7 years (A)
P14	Jan Smole	*(formerly)* Trust & Safety Process Manager, TikTok *(formerly)* Community Operations, Facebook	13 years (I)
P15	Kristen Vaccaro	Assistant Professor, University of California San Diego *(formerly)* Researcher, University of Illinois Urbana-Champaign	8 years (A)
P16	Sarah Myers West	Managing Director, AI Now Institute Senior Advisor, Federal Trade Commission	8 years (C,G)

Table 5.2: Anonymized participants (self-described), including academic (A), other civil society (C), government (G), technology industry (I), and moderator (M) experiences at time of data collection (2022).

	Role(s)	**Experience**
P17	Associate Professor Moderator	5 years (A,M)
P18	Lecturer, Northeastern University Researcher, Stanford University Moderator	2 years (A,M)
P19	Researcher Moderator	5 years (A,M)
P20	Research Scientist	3 years (A)
P21	Researcher, Twitter *(formerly)* Researcher, Facebook	6 years (I)
P22	Researcher Program Manager	4 years (A,C)
P23	Researcher	5 years (A,I)
P24	Researcher *(formerly)* Researcher, Meta	3 years (I)
P25	Research Manager, Meta Researcher, Meta	4 years (I)
P26	Senior Designer, Game Company	10 years (I)
P27	Data Scientist, Google *(formerly)* Data Scientist, Facebook	7 years (I)
P28	Data Science Manager, Twitter Data Scientist, Twitter	2 years (I)
P29	Product Manager, Meta *(formerly)* Product Manager, Microsoft	4 years (I)
P30	Policy Manager *(formerly)* Trust & Safety Project Manager, Facebook *(formerly)* Market Specialist, Facebook	4.5 years (I)
P31	Program Manager, Community Safety, Discord *(formerly)* Trust & Safety Specialist, Lyft *(formerly)* Trust & Safety Specialist, Discord *(formerly)* Trust & Safety Specialist, 24/7 InTouch at Airbnb	7.5 years (I,M)
P32	Journalist, Washington Post	5 years (C)
P33	Data Scientist, Twitter *(formerly)* Researcher, Facebook	5+ years (I)

To facilitate this collaborative problem-solving, workshops began (following a brief introduction) with a virtual adaptation of the KJ method of brainstorming (Kawakita et al., 1967; Scupin, 1997). Participants were asked to independently identify obstacles to successfully moderating content at scale, using Google Jamboard (a virtual brainstorming tool). After collaboratively grouping similar ideas under larger themes, each participant noted their top three priorities, or the most important obstacles to resolve, resulting in a categorized set of collective priorities. Next, participants engaged in a paired journey mapping exercise to more tangibly illustrate their understanding of current scaled content moderation processes. After each pair shared their journey map with the larger group, participants were asked to reimagine their maps using the magic wand approach, temporarily setting aside any practical obstacles (e.g., resource constraints). Finally, participants collaboratively generated a roadmap to move us closer from the status quo to their collectively imagined future, populating a feasibility grid with key objectives—i.e., what can be accomplished in the next year? 5 years? 10 years?—with a focus on specific tasks and outcomes.

I conducted a thematic analysis (Braun & Clarke, 2006) of the resulting data, which included voice and chat transcripts as well as visual artifacts collaboratively produced by participants during each workshop. I used an inductive approach to develop codes, individually reading workshop transcripts and noting codes by hand. After discussing these initial codes with a research assistant, I created a more comprehensive list of codes (36 codes in total). I manually coded two transcripts in a pilot coding process to test and refine the codebook. Resulting codes were organized around several themes, including but not limited to cultural context; issues of scale; and corporate values and practices. Quotations have been lightly edited for readability.

I also draw from methods of situational analysis (Clarke, 2005), utilizing social world mapping (see Figure 5.1) to visualize the influences of and relationships between the collective actors involved in negotiating contemporary social platform governance. Situational analysis provides a framework for interpreting how different groups engage in complex negotiation, facilitating analysis of competing interests and power differences. My analysis draws on multiple intersecting data sources, including workshop transcripts, visual artifacts, and other discursive materials, as well as my own experiences in the professional content moderation industry.

5.4.1.1 Position statement

My research questions, analysis, and contributions cannot be understood independently of my position as the researcher, which includes my experiences as a white, cisgender woman living in the United States (Williams & Irani, 2010; Bardzell & Bardzell, 2011). This particular study and its methodological approach are made possible by my unique position as both an academic researcher and an industry practitioner, which affords me a unique level of familiarity with the inner workings of major technology companies. In addition to being a PhD candidate at the University of Michigan (where I have conducted social media research since 2014), I have been directly employed by several different technology companies, beginning with Meta (then Facebook) in 2017. I have also been employed by Twitter (now X) and two social media startups (Sidechat and YikYak), and I currently lead Trust & Safety operations at Mozilla. While the present research is not an ethnographic inquiry, it is necessarily informed by years of professional research about, and applied practice in, social media governance. My research is also informed by my experiences as a queer, disabled person and technology worker, and by my personal experiences with online harassment and other abusive behaviors technology companies typically wish to prevent on their platforms.

5.4.1.2 Limitations

This research represents perspectives from participants who are largely (though not exclusively) Western, white, and highly educated, reflecting broader trends in the technology workforce (U.S. Equal Employment Opportunity Commission, 2024). Experiential understanding from more precarious technology governance workers is also notably absent from the current work, due to the risks associated with study participation as well as the limitations of my own professional network. Future research should investigate these additional perspectives, including those of workers located in the Majority World.

5.5 Results

The results are organized into three sections: content moderation foundations (*what are we trying to achieve?*), content moderation realities (*where are we falling short?*), and content moderation futures (*what can be done?*).

Content moderation foundations. What are we trying to achieve? Across social worlds, participants agreed that successful content moderation is *principled, consistent, contextual,* *proactive, transparent,* and *accountable.* First and most critically, participants emphasized the impossibility of "neutral" governance, urging platforms to commit to specific values instead of striving for the unattainable goal of universal applicability. Efforts to appear impartial yield vague, imprecise policies that cannot be operationalized at scale—but like other forms of governance, successful moderation requires rules to be consistently enforced, in order to establish clear expectations for appropriate behavior. Successful moderation also accounts for relevant context, though this is difficult (if not impossible) to achieve at scale, given both the volume of content users produce and its global context. Given the challenges associated with scaling reactive moderation, participants advocated for more proactive—and preventative—approaches to mitigating harm. Finally, participants stressed that successful moderation is both transparent and accountable: users must understand how and why decisions are made, with meaningful opportunities for recourse when mistakes inevitably occur. Transparency and accountability not only build trust but also confer legitimacy, encouraging compliance even when users disagree with individual policies or outcomes.

Content moderation realities. Where are we falling short? Participants described numerous barriers to achieving these goals on contemporary social media platforms, including the impractical pursuit of "one size fits all" solutions to global governance—resulting in vague, Western-centric policies that ignore cultural nuance and further marginalize non-dominant groups. These structural flaws are compounded by the demands of scale, as machine learning models can only reliably detect content with broad agreement. Despite these challenges, many technology companies pursue what participants described as a "growth at all costs" business model, creating strong incentives that frequently and directly conflict with user safety goals. Participants characterized company leaders as reluctant to invest in content moderation and other work related to user safety, often prioritizing new product development over the mitigation of existing product risks. Despite limited resources, workers are expected to support continual platform expansion and scaled growth, resulting in what participants described as an overreliance on third-party labor and other exploitative working arrangements. Content moderation labor and other data labeling tasks are outsourced to vendor workforces in regions with lower labor costs, while civil society organizations engage in unpaid labor they may never see implemented. Individual workers are left feeling unimportant, overburdened, and isolated from their closest peers, resulting in what participants describe as chronic burnout and frequent employee turnover.

Table 5.3: Overview of results.

Content moderation foundations: What are we trying to achieve?		
Successful moderation is **principled**	*Platforms must define their values; neutrality is impossible*	p. 86
Successful moderation is **consistent**	*Consistent enforcement creates clear expectations*	p. 87
Successful moderation is **contextual**	*Accurate, equitable outcomes consider relevant context*	p. 88
Successful moderation is **proactive**	*Instead of reacting to harm, prevent harm from occurring*	p. 88
Successful moderation is **transparent**	*Users should understand how and why decisions are made*	p. 89
Successful moderation is **accountable**	*Avenues for recourse cultivate trust, process legitimacy*	p. 91
Content moderation realities: Where are we falling short?		
One size fits none: The futile pursuit of universal agreement	*Lack of industry consensus restricts cohesive progress*	p. 92
	When global rules are defined locally, Western perspectives dominate	p. 93
	Striving for universal applicability dilutes policy effectiveness	p. 94
Problems of scale: "Lowest common denominator" moderation	*Crude solutions can't be scaled*	p. 95
	Scale flattens nuance	p. 96
Growth at all costs: Misalignment with business incentives	*Business incentives undermine safety goals*	p. 97
	Metrics obsession drives optimization of numbers, not outcomes	p. 98
	Flawed "innovations" prioritized over fundamental safety	p. 99
	Preoccupation with short-term costs prevents long-term gains	p. 100
	Constant deprioritization results in chronic underinvestment	p. 102
Labor pains: Exploitation, isolation, and burnout	*Overreliance on third parties and exploitative working arrangements*	p. 103
	Fractured understanding: Navigating organizational silos	p. 104
	Burnout, turnover, and institutional knowledge loss	p. 106
	Executives enjoy disproportionate influence	p. 107
Content moderation futures: What can be done?		
Embrace **holistic visions**	*Moderation is essential, forever, and communal*	p. 109
Design platforms that **encourage prosocial behavior**	*Establish clear expectations for appropriate behavior*	p. 112
	Promote rehabilitation, not retribution	p. 114
	Favor flexible interventions	p. 115
	Implement responsive regulation	p. 116
Incentivize **corporate accountability**	*Regulation: Promising, but not a panacea*	p. 118
	Leverage alternative accountability mechanisms	p. 120
Explore **alternative models**	*Forsake the 'killer app'*	p. 122
	Empower community governance	p. 123

Content moderation futures. What can be done? After identifying numerous ways in which contemporary technology companies fail to successfully govern their platforms, participants considered potential solutions. First, participants encouraged company leaders to clearly articulate a holistic vision of community safety—one that recognizes content moderation as an integral component of the product experience, rather than a crisis response. Participants highlighted the role of platform design in shaping user behavior, advocating for designs that establish clear expectations for appropriate conduct and promote the rehabilitation of potential offenders. Because existing corporate structures have failed to adequately protect users' safety, participants stressed the need for greater corporate accountability, whether incentivized via regulatory structures or alternative levers. While participants expressed support for the development of new regulations to force more responsible corporate action, they questioned lawmakers' proficiency in a problem space rife with complexity and nuance—and they cautioned that technology companies may be incentivized to identify loopholes to avoid increased scrutiny. Finally, given the challenges associated with uniform governance of general-purpose platforms, participants advocated for more localized and participatory alternatives, empowering users to shape and steward their own community spaces.

5.5.1 Content moderation foundations: What are we trying to achieve?

A foundational challenge identified by many participants is the lack of a unified, strategic vision for what content moderation should ideally achieve, particularly at the scale afforded by modern social media platforms. P9 said:

> "Obviously, one of the biggest challenges is that there actually isn't a real definition of what successfully moderating social media at scale looks like in the first place. What is 'success' in that context? What does it mean for it to be successful?"

5.5.1.1 Successful moderation is principled

First and most critically, participants noted the importance of establishing foundational principles to guide moderation practices. "User perceptions of harm can be complex," said P23. "But in a content moderation context, I think platforms have to take a stand. Basically, it becomes a question of: are you going to take the user-centered approach? What's the cost that comes with that? If there's one group of users who think child sexual abuse material is completely okay, are you going to listen to them?"

A failure to establish consistent values—often in favor of something resembling "neutrality"—can instead reinforce existing systems of social oppression, such as racism and transphobia. P24 said: "Some criticism I've heard, from within and outside of the company, is that a lot of the policies are written in a perspective that favors certain groups. There's definitely been a lot of criticism of how people of color are treated, how LGBTQ+ people are treated, how overweight or plus-size people are treated, and how the policy tends to favor cis white folks." P7 agreed:

> "You see that a lot with content creators. I follow streamers who are activists as well, and some of them have actually had conversations with companies like Twitter and been like 'Yes, I was in fact shadow-banned, because I was being loud on main.' And then you see folks talking about 'I can't swear, I can't talk about being black, I can't talk about being gay, I can't talk about this, that, or the other thing,' which is a problem."

Instead, many participants emphasized that neutrality is not an achievable goal. P32 said: "There is no global set of values. At some point, you just stop trying to please everybody and make a choice. What are *your* values?"

5.5.1.2 Successful moderation is consistent

Attempts to remain impartial can produce ambiguous policies that are difficult or even impossible to successfully operationalize at scale, resulting in inconsistent and imprecise moderation. Successful moderation, like other forms of governance, requires rules to be consistently applied. "Moderation needs to be active and continuous in order for internal norms to be maintained," said P27. Inconsistent moderation—whether due to middling principles or unreliable enforcement—results in unclear expectations for user behavior. P27 continued:

> "If we don't have clarity around what the rules are within a particular community, people will be more inclined to break those rules due to lack of knowledge or lack of clarity as to what is considered acceptable."

Inconsistent moderation also creates additional opportunities for malicious actors, who may be willing to risk potential enforcement when consequences are not guaranteed. "Consistent response is important," said P33. "Inconsistent, slow enforcement creates space for adversarial users to think, ‘It’s worth it.’ A scammer with three days to scam someone is likely to think it’s worth the effort.”

5.5.1.3 Successful moderation is contextual

Effective moderation also considers context, which becomes increasingly challenging at scale, given both the volume of content users produce and its global context. For individual moderators, additional time spent reviewing particularly challenging or otherwise ambiguous content may jeopardize the metrics by which their performance is strictly evaluated—further complicated by standardized review tools that may not display all potentially relevant information. “When you are on the ground doing this kind of evaluation, it can be really challenging,” said P6. “Even if you have a really solid policy, it can be challenging to make correct determinations based on that, because there's so much missing context—context that you might not actually have access to. It also takes time, and you have to respond to some of these issues really, really quickly.”

Even without the constraints of time and tooling, context can be difficult to glean; language evolves quickly, and online interactions occur in diverse social contexts. “A lot of harms are referenced by coded language—and the coded language can evolve faster than people can detect it,” said P8. “Also, so much of this is situated within the context of friendship. It's sometimes hard to tell the difference between friendly banter and true harm.” While it isn’t realistic to expect all relevant information to always be available (or accessible), providing moderators with as much context as possible is critical to producing accurate and equitable governance outcomes.

5.5.1.4 Successful moderation is proactive

Many participants highlighted the industry’s focus on reactive moderation practices, which P9 described as “an over-focus on interventions being at the point where harm has already been done in some way—versus attempting to intervene potentially before that point.” P10 agreed:

> “Most moderation is happening far too late. More proactive and upstream work is rarely prioritized or thought about. Often we're talking about the decision that happened, but not the days, months, weeks, years that led up to that point in time.”

Participants strongly favored more proactive strategies, such as improved user education and other risk mitigation efforts designed to prevent at least some policy-violating content from ever appearing on the platform. P7 said:

> “I've worked in operations for a very long time. We want to react; we don't proactively create systems that can address some of the problems. You can't catch everything ahead of time—there is no way to predict a violation before it happens—but you can have systems in place. Twitter has the thing where it's like, ‘Are you sure you want to send that?’ Interference and friction is a good way to stop things from happening.”

Part of this over-reliance on reactive moderation is a function of underinvestment, with many companies only turning their attention and resources toward platform governance once some issue has already arisen. P5 said:

> “I have seen with a lot of start-ups something you might call the ‘Oh Shit’ model, where very little if anything is ever getting moderated. The first real moderation actions come in reaction to some personal referral. Somebody knows somebody at the company, an engineer or PM or executive... I’ve seen this happen in at least two start-ups, and I wouldn’t be surprised if it’s most start-ups.”

To the extent that platforms do invest in proactive mitigations, it is typically natural language processing algorithms designed to detect potentially violative content before a user reports it. “‘Proactive’ in industry tends to mean getting rid of something before it's reported,” said P13. “To me, proactive means something much more like preventative care in healthcare—the issue was prevented.”

5.5.1.5 Successful moderation is transparent

Participants widely agreed that successful moderation is transparent, and that a lack of transparency erodes user trust. P27 said:

> “When active moderation takes place, if there is a lack of transparency in the decision-making process, that can lead to frustration and ferment dissatisfaction and fracturing in the community. A successful community is one where when hard decisions are made,

> there is clarity and transparency around the final outcome. Any successful community needs to have a transparent moderation process in place."

Several participants referenced platforms' belief that increased transparency will lead to undesirable outcomes, as "bad actors" may be better positioned to manipulate their systems. "Companies are reluctant to be transparent about their moderation processes because they are afraid of the bad actors who game their systems," said P20. Instead, many platforms restrict what they share with the public, often obscuring important details and fracturing users' understanding of what is or is not allowed. P20 continued: "The community rules are usually vague, so different users have different interpretations of the rules."

In the absence of meaningful transparency, users are left confused by moderation decisions they do encounter on the platform; they may not even recognize that moderation has occurred. "I don't think content moderation can ever be called successful when most users have no knowledge of it," said P1. "And no agency over it. At all." P4 agreed, suggesting that users who do engage in policy-violating behavior are "not being told what the problem is":

> "TikTok's moderation system has some problems. One of them is a severe lack of transparency when it comes to what rule has been violated—why someone's content is being removed, or why someone has been banned. I've been told by other creators that when you hit the appeal button, there used to be a 'Tell us why you're appealing.' Now, there's not. It's just, 'This content was removed because of a Community Guidelines violation.' You click a button to appeal, and what happens after that, you don't know. At some point in the future, you either get something back that says 'We reviewed your content, and it was restored'—and you still don't know what they think you did wrong—or 'We reviewed your content and decided it was a violation,' and you don't know what you did wrong and whatever rule you broke. You don't know how to not break it again."

Reduced transparency also limits the extent to which content moderation experts and the general public can provide feedback to platforms, hindering potential improvements to their moderation systems and practices. "This lack of defensible and extremely transparent governance practices restricts the ability for companies to receive feedback," P25 said, noting that platforms should publicly outline "how our systems view and handle each type of problematic content and solicit feedback from experts on improving the status quo."

5.5.1.6 Successful moderation is accountable

Finally, participants agree that successful moderation practices must have mechanisms for ensuring accountability, both to users and the broader public. Accountability is critical not only for external purposes (such as meeting evolving regulatory requirements), but because fully accurate content moderation is simply not possible, particularly at scale. When mistakes inevitably occur, accountability mechanisms such as appeals processes afford users some amount of recourse. P4 said:

> "Errors and automated moderation… the combination of these two things is really bad. Like, if you're going to have errors—which, I mean, you're going to—especially if they might be systemically biased, which we also see some evidence of on TikTok, then you have to have a good appeals process."

As with offline systems of governance, accountability is paramount to ensuring user trust. "A big obstacle to successfully moderating at scale? Process legitimacy," said P17. "You could have the most effective triage process, but if people don't believe in it producing fair outcomes, then they're just going to undermine it. Looking at recent Supreme Court decisions, we see that playing out in different ways." P6 agreed that accountability is critical in building trust between a platform and its users—or, conversely, that a lack of accountability will have the opposite effect:

> "Building trust in the system, between users, and the policies, and how they're implemented… that is especially important now, where we've had a number of scandals—across a number of different platforms—that have all resulted in eroding trust between users and platforms."

Participants also emphasized the distinction between governance legitimacy and endorsement, noting that even users who disagree with individual policies may still perceive them as legitimate if the process of creating and enforcing those policies is believed to be fair. "People don't have to all agree with all the decisions," said P29. "But they have to respect the process by which we made them, and they have to respect where we have chosen to draw the line." P26 agreed: "I think process legitimacy, depending on the circumstance, can be really important—but as long as you're in that window of legitimacy, it doesn't have to be super well-received. It just has to be something that people can't really push back on." This perspective aligns well with *procedural justice* (Tyler, 2006), which suggests that people are more willing to comply with rules when they perceive the organizations responsible for their enforcement as legitimate authorities (Tyler, 2007; Katsaros et al., 2022). P29 continued:

> "If you can sort of get to the point where, through the combination of process and outcome, you have something that is sufficiently defensible, then hopefully, you can get at least the good faith actors in this ecosystem—which, granted, a lot of the people who yell very loudly are not—to the point where they understand the decisions, respect the decisions, are willing to follow the decisions… even if they're not necessarily willing to say, 'Yes, this is something I would stand behind or agree with.'"

In the absence of global consensus, process legitimacy functions as a stabilizing force—but as participants emphasized, platforms must deliberately cultivate institutional trust to maintain governance authority. "Efforts to affect the *efficacy* of moderation are often orthogonal to efforts to affect *perception* of moderation," P26 continued. "If you make your detection better, oftentimes that doesn't pay it forward in terms of people thinking that you're making better decisions. You have to work on both."

5.5.2 Content moderation realities: Where are we falling short?

While participants agreed on what characterizes successful moderation at scale, they also described numerous ways in which contemporary platforms are falling short of stated goals.

5.5.2.1 One size fits none: The futile pursuit of universal agreement

While participants overwhelmingly agree that successful content moderation is both principled and consistent, most participants characterized current moderation practices as inconsistent at best—in large part due to the difficulty (if not impossibility) of creating policies, protocols, and other governance mechanisms designed for global application.

Lack of industry consensus restricts cohesive progress. Participants expressed frustration with what they perceived as a lack of industry standards, preventing consistent progress across company lines. "There's no general consensus among experts or users on what is an adequate system," said P25. "The conversation is always evolving. What is overreaching? What is under-supporting users?" Even key terms used across the industry, such as *online harassment*, are not consistently defined. "How feminists define harassment can be very different than the people using #Gamergate," said P4. "I always remember a quote from a paper where they interviewed people who identified as part of Gamergate. They were like, 'Oh, that's not harassment. That's just words.'"

For many participants, this lack of standardization across the content moderation industry—or even within a single company—results from the inappropriately monolithic treatment of large, global userbases. "There's this vision of a 'free speech town square' for big social media platforms, versus a community with agreed norms," said P14. "I think it's a core issue. As much as these companies stress 'community' in public statements, there isn't really a community, or any agreed norms for that community." P9 agreed: "Millions of people are a bit too big to just be one community, with shared values and beliefs and principles and the like. Treating all the different communities that exist and interact with each other as somehow one giant community almost inevitably leads to pain."

When global rules are defined locally, Western perspectives dominate. In the absence of global standards, companies must define their own—resulting in governance policies and processes dominated by largely Western perspectives. "Part of what makes a lot of the cases we see in this space much worse is when our systems are imagined by and encode a very Western, cisgender, white, straight perspective into the decisions that are being made," said P26. P7 agreed: "If you have a room full of cishet white men who are all 50 and up, you have a problem. Having centralized policies and guidelines for moderation is very important—but they're only very good and very helpful if you have a lot of voices in the room, all being heard."

P32 referenced this "lack of cultural awareness" as a major obstacle to successful scaled moderation. "The majority of policies are written by people in the United States—and they apply to people all over the world," P32 said. "They're not even in their languages. Considering that most users are actually outside the United States, it's just a fundamentally broken piece of the system." P21 agreed:

> "There's a lack of cultural context regarding harassment, abuse, intimate imagery, toxicity, privacy… what is shame? What is embarrassment? I've done a lot of work in non-Western countries, and what privacy or intimate imagery is in Saudi Arabia is not

> what it is in Finland. When it comes to content moderation, we really run into a lot of problems because we can't have country- or culture-specific policies."

Participants emphasized that content policies, moderation processes, and other product experiences do not adequately account for the experiences of marginalized people. "I come from the activist side," said P22. "What we're seeing is an interest in viewing the experiences of marginalized groups—people who are not necessarily, you know, privileged users of whatever tech platform—as edge cases, rather than the baseline. That's a huge issue, especially when we think about refugees, trans women, LGBTQ individuals in general, women's rights… all of these things. We are always kept on the margin, and dealt with—for policy application as well as remediation—as edge cases, which creates a lot of issues. At the end of the day, there is no context applied, even though a lot of tech companies say that they engage with stakeholders from different regions."

Striving for universal applicability dilutes policy effectiveness. Though some amount of standardization is required to progress scaled solutions, participants emphasized that universal agreement simply is not possible. "You are looking for agreement on a taxonomy," said P15. "What is toxic? What should be removed? What you're asking for is a world without humans. No matter how toxic something is, there will still be some disagreement. And as soon as you get something that's more 'borderline,' there's going to be a whole lot of disagreement."

While companies could choose to adopt a more principled stance, they hope to attract the broadest possible customer base—resulting in ineffective, overly generalized policies that lack the specificity and coherence to be effectively implemented at scale. "Policies are unclear, ever-changing, reactive, and overwhelming in amount and detail," said P14. "This super granular policy that dances around the issues, rather than a proper set of norms that the community agreed to by joining, is a massive problem—with a lot of problems coming downstream from that." P17 agreed: "Different identity groups will have very different points of view on what toes versus crosses the line. Platforms end up making these very decontextualized policies that apply across many different application areas, and thus really piss a lot of people off."

Ultimately, participants recognized necessary trade-offs between content moderation practices that more appropriately account for individual differences and those that efficiently scale. P1 said:

> “The dilemma at the heart of this is that modern moderation at scale requires some universal agreement—but we also want to honor all of the cultural differences and other differences among people, and that seems to make doing anything at scale especially impossible. You could establish universal definitions—as we have for lots of things, especially in international law—and then do moderation differently in different cultures and different contexts, in order to respect those different cultures and contexts and the values of those users. But, as somebody pointed out earlier, sometimes users have what we consider terrible ideas. So it's extremely tough, you know, balancing opposing input and essentially imposing the ideas of, let's face it, over-educated cultural elites from a few countries on lots of other people—the global majority.”

5.5.2.2 Problems of scale: “Lowest common denominator” moderation

Despite this lack of standardization, content moderation processes must still be scaled—resulting in crude solutions that predominantly address issues with broad consensus, at the expense of more nuanced concerns.

Crude solutions can’t be scaled. While many companies rely on machine learning models to automatically detect potentially violative content, the performance of these models is dependent on the quality of the data used to train them. Accurate detection, for example, relies on clear, consistent definitions—the very consensus still lacking in many safety-adjacent problem areas. “From some of our own work, we know that in one of the biggest datasets, even the stuff that’s rated as ‘toxic’ has like a third of annotators disagreeing,” said P17. “It’s just a very highly disagreed upon area.” P26 agreed: “How do we label something as ‘toxic?’ That can be super subjective.”

Participants reflected on the inherent limitations of attempting to automate detection of content that even humans cannot consistently identify. “There's a really interesting paper by Mitchell Gordon called ‘The Disagreement Deconvolution,’” said P13. “He wrote it for a machine learning audience, but he's basically quantifying the argument that reasonable people disagree. There's a kind of peak to model performance where we can't get better—because there isn't objective truth above a certain level.” Even for more straightforward types of content, any classification task is subject to certain limitations, such as the inverse relationship between *precision* (e.g., of all emails marked as spam, how many actually contained spam?) and *recall* (e.g., of all actual spam emails received, how many were detected?). P28 noted that optimizing precision and recall is only more challenging when attempting to automate content detection in complex problem areas:

> "When we speak about moderation at scale, the easiest way to scale—and the cheapest way—is through algorithms and AI. But all algorithms, all machines, have precision-recall tradeoffs: they all have some false positives and false negatives. And the more challenging a problem space is, the higher chance that precision-recall tradeoff will be suboptimal—it's really hard to achieve good precision-recall in a difficult space. So those most difficult cases end up manually moderated, because algorithms are not detecting them properly. And now a human needs to make a judgment, whether the content is in violation of policies—but again, this is the most difficult type of content, so there is a lot of subjectivity and interpretation of guidelines. AI cannot really help with this. This is where it reaches its prediction power plateau."

Scale flattens nuance. Several participants noted the ways in which scale inherently minimizes difference, resulting in moderation practices that cannot appropriately consider context. "Automated tools flatten behaviors," said P19. "They get more abstract, and then you start to miss local context."

Participants emphasized the difficulty of consistently enforcing policies across communities with divergent norms. "It's very hard for decisions to strike the right balance between nuance and consistency," said P29. "The inherent tension is that, for a really clear process, you want something that's relatively clear-cut—but then it runs up against, like, real life, and the fact that communities do have different norms." P26 agreed:

> "One of the challenges I see is folks sort of colliding in spaces, with very low context for each other and coming from different groups, or even just different circumstances—like 'Hey, I'm having a bad day; I'm not resilient to somebody being mean to me.' Simplifying that basically means that the average, the default, the lowest common denominator moderation does not set anybody up for success in those interactions, however you define success. And as a result, a lot of our systems either flatten or discourage expression of different identities—or make it so that spaces are differently friendly to particular cultures or groups. Sometimes that's desirable; sometimes it's not."

Instead of reckoning with this complexity, many platforms focus their attention on moderating less nuanced violations. “Most platforms wind up with very, very weak table stakes because they don't like the fact that there's all of these diverse perspectives and disagreements,” said P17. “We end up only moderating the stuff that is sort of universally agreed upon—like, a racial epithet, that should definitely go. The stuff that is more in the gray area tends to not be moderated, or left up to the communities themselves.”

When complex human interactions are evaluated in accordance with decontextualized rules designed for scaled enforcement, predominant norms prevail—often reproducing global power asymmetries. P6 reflected on the reinforcement of dominant experiences at scale:

> “In the community that I moderate, sometimes we don’t know. Like, the use of the C-word—totally unacceptable in the U.S., totally normal in Australia. So what do you do when somebody is using that? Do you just reinforce an already American hegemony of platforms and technological spaces by getting rid of that? Or do you potentially have this word that is interpreted by a lot of people as misogynistic? There’s all these trade-offs that have to do with power that are really challenging to scale.”

Participants noted the ways in which this ‘flattening’ of individual expression risks further marginalizing vulnerable users. “Even when we sit down and go back and forth on like… okay, yes, in many contexts, this term is hate speech, it’s more complicated than that for some groups,” said P26. “And when our systems flatten a decision across all groups, we need to be really thoughtful about why—and what the consequences of that are.”

5.5.2.3 Growth at all costs: Misalignment with business incentives

Despite the challenges inherent to global scale, many contemporary businesses pursue what P5 described as a “growth at all costs” model, designed to increase profits and maximize shareholder value—creating strong incentives that frequently and directly conflict with typical Trust & Safety goals.

Business incentives undermine safety goals. When asked about the most significant barriers to successful content moderation at scale, P11 said: “Shareholders, venture capital… I'm trying to stop short of saying ‘capitalism.’” P5 agreed:

> "One way to look at the entire problem is as a negative externality of venture capital—and the venture capital model of funding platforms. You know, the 'growth at all costs' incentives that platforms have at various points in their lifecycles. I've talked with lots of students who have worked at various companies, and leaders at different platforms, and the folks I talk to always describe this process as an evolving nuisance to leadership. So I think there are some pretty big structural impediments to actually addressing this at scale."

These market pressures manifest in company-level goals, often requiring individual teams to prioritize user growth over user safety. "Company-level metrics like growth and engagement are in opposition with content moderation," said P3. "At Facebook, anything introduced to reduce harm also reduces engagement. You are running up against a top-line metric for the company that impacts the stock price—that everyone across the board is optimizing for."

Participants noted that content moderation efforts are often treated as compliance checkboxes, rather than sincere attempts to reduce harm. "A lot of the current goals of moderation are about compliance—making sure we don't have the bad content, because we'll get in trouble," said P4. "What would it look like instead for that to be about 'Are we actually hurting people?' I'm not saying platforms don't care about this at all, but a lot of things do come down to business incentives. If companies didn't have to make money, a lot of things would be really different." P3 agreed, drawing a parallel to for-profit healthcare:

> "It's very hard for us to incentivize ourselves to have the right outcomes in medicine if things are for profit. I think the same is true for social media companies. In a world in which we are trying to profitize interactions between people through 'growth' and 'engagement,' we'll never be able to solve some of these problems. If you're expecting a pharmaceutical company to grow year-over-year, that means more people need to be taking these drugs—which means you're solving less of the underlying problems, and just hoping to continue treating the symptoms, more and more every year."

Metrics obsession drives optimization of numbers, not outcomes. While measurement is essential to understanding performance at scale, an overreliance on quantification and optimization can distort incentives, undermining broader goals. "When I'm trying to implement safety best practices, I often get pushback like, ‘We basically only care about numbers,’” said P31. “They're very honest about that. Balancing different goals from people who are not safety-focused... it's so hard sometimes. It's really difficult.”

Participants stressed the difficulty of developing meaningful metrics to assess complex and highly contextual harms—further widening the divide between Trust & Safety and business goals. “It's easier to report total takedowns, rather than—for example—total problems prevented by proactive interventions,” said P13. P26 agreed:

> “Basically no one feels like they have good KPIs or measurements for success in this space. We're all really struggling… especially convincing the rest of your org that you are doing well. It doesn't mean that we're not trying, but it's an ongoing struggle.”

Even where reliable, meaningful measurement is possible, user safety metrics are often in conflict with broader organizational goals. “It seems impossible to make the right decisions from a harm reduction standpoint if your goal is engagement, because harmful content is engaging,” said P3. “It catches your eye—you want to read it; you want to see what’s going on. A lot of times, that makes it really hard for these companies to actually take the hard step to reduce harm.” Participants expressed similar concerns about common operational metrics, such as *average handle time* (i.e., the average duration of a content moderation decision), which encourage content moderators to prioritize speed over quality. “Productivity and accuracy metrics are one way of measuring, but it's probably not the best way,” said P12. “That setup can lead to different types of shortcuts that eventually cause flaws in the system.” P14 agreed:

> “They’re under high productivity pressure—and what they do is game the system. They find ways to cut corners and make do with those productivity targets, which then leads to mistakes.”

Flawed “innovations” prioritized over fundamental safety. Participants felt that the rapid pace of product development within the technology sector prevents adequate risk mitigation, with innovation taking precedence over necessary product maintenance—including solutions to existing problems. P3 reflected on the difficulty of prioritizing foundational work in environments that incentivize novelty:

> "At Facebook, you're pretty much in a grind on the performance review cycle. A lot of times, the sort of like, baseline work isn't valued as much as flashy projects that get a lot of attention from a leadership level—so people drop a lot of the really hard but essential work to do content moderation well at scale. Things like working with the outsourcing sites. It's grueling work. It's very time-consuming. It requires a lot of time and energy. But that's not the work that's valued when you go through a performance review cycle—what's valued is a new feature launch or a large engineering project; something new and flashy. No one wants to do that work, and the people who do generally aren't rewarded for it in the right way."

Participants expressed particular frustration when companies shift focus to new product areas before fully addressing known problems in existing products. "I think another big issue is companies prioritizing or adding new areas of focus," said P24. "For example, how Facebook became Meta. There's this new huge focus on VR. It's still kind of crazy to me that there is this huge shift in focus when they haven't even solved the problems on the regular platforms." "There are constantly new products that we need to be supporting," added P8.

Participants emphasized the importance of comprehensive testing prior to a product's deployment to large populations of users. P8 expressed concerns about the rapid development of machine learning models, which he felt are not well understood even by those who build and maintain them:

> "One of the other challenges is, as we start using a lot more machine learning models, no one really understands how these models interact at scale. I keep hearing people say 'you need to release the algorithm.' There is no algorithm—there are hundreds of algorithms, and they all have emergent properties that we don't understand, even inside the company. Why was this decision taken? No one can explain that."

Preoccupation with short-term costs prevents long-term gains. Participants expressed frustration about the relentless pursuit of maximized earnings, particularly given regular spending reductions in areas deemed non-essential by company executives. "I don't really have an answer to the capitalism thing," said P19. "They're trying to run a large social media platform. There's always people who work in operations trying to make sure they're cutting costs in ways that can bring revenue into the company." Industry participants described company leaders as reluctant to allocate resources to initiatives that impact potential profits less directly. P28 said:

> "There's so much pushback, because it's like, 'But that costs money to change right now. I could spend the money later.' Yeah, but you could also know a thing is coming. If you know a thing is going to be better long-term, there needs to be better incentives to doing that thing, like harm reduction in moderation. If you can find a way to show where this harm reduction will, you know, not necessarily create better profits, but create better user experience—so that you have better user retention, so that we have better turnover rates—that's important to show, because then you can incentivize these things."

While the costs associated with scaled content moderation are immediate and easily observable, the benefits are both less explicit and cumulative, occurring over time—and only when moderation is successful. Moderation mistakes were characterized by participants as both inevitable and costly. "When you have something that has been moderated but maybe shouldn't have been—or something that maybe should have been moderated, but wasn't—the news of that can travel pretty quickly," said P19. "There are cascading effects of decisions that are done poorly, or just by the basic error of whoever's making the decision, that make it challenging to maintain trust in the governance of the platform." P27 agreed, emphasizing the fundamental imbalance between the longer-term shifts in community norms successful moderation can enable and the immediate costs of moderation mistakes:

> "There's an asymmetry inherent in scaled content moderation: if you moderate badly, the observed downsides are immediately apparent, in that people will see incorrect decisions being made. It leads to lack of trust in a pretty short-term time cycle—versus failing to moderate at all, the community can kind of maintain itself in the short term. The downsides in being hands-off only start to become apparent longer-term, and are much harder to quantify. That asymmetry leads to an implicit bias, where we tend to do nothing because it's easier to justify, at least in the short term. I think that's a very difficult and profound obstacle to overcome. There's always this subtle nudge to doing less than doing more, because we are more easily able to quantify the downsides of doing more."

Constant deprioritization results in chronic underinvestment. This misalignment with foundational business incentives results in what industry participants described as chronic underinvestment in content moderation and other work related to user safety. When asked about what could be done to improve the current state of scaled content moderation, P24 said: "Just investing more in integrity at these companies. This is a high-level value to users. We already are investing, but definitely not enough." P31 agreed: "Not enough money or employees focused on content moderation, combined with greed—tech companies and their motivation for existing, and motivation for scaling, how they prioritize growing as a company and where they dedicate their internal resources." P16 described challenges with even "basic resources, like having enough time to make evaluations."

Other participants questioned whether additional resources could sufficiently propel improvement without other, more systemic changes. "Something Amy Bruckman says in her book is that if we had unlimited resources, content moderation would be perfect—and I'm not sure I completely agree with that," said P4. "Some of the challenges here could not be solved by resources; things like values conflicts. In theory, a platform could hire enough people to look at every single piece of content within seconds that it's posted, and also pay them well and train them well and that sort of thing. Of course, that would be a massive amount of resources—and again, it wouldn't solve all of the problems." P10 agreed: "Facebook is a counterexample. Facebook has spent… it's a crazy amount of money. It's in the billions. They have tens of thousands of people; they've put in a lot of resources. But I don't know if the problem has gotten any better on Facebook. Some might say it's gotten worse." "And we're also constantly still hearing about poor conditions for the labor force," added P4. "So there could be more resources going into that, whether it's paying people more or mental health support."

Ultimately, participants overwhelmingly agreed that technology companies could and should be investing more in content moderation and other efforts to protect the safety of their users, highlighting the significant costs of underinvestment. P31 said:

> "If a company is not dedicating enough resources to cover all of its content moderation needs, the result of that, potentially, is massive widespread harm to society. We've seen how angry society is these days. Incels, violent extremism, all sorts of stuff online is causing real-world harm. Facebook has even come out with data that shows they knew they were damaging the mental health of teenage women on their platform. It's just really

> upsetting to think about how many companies are kind of intentionally sacrificing the well-being of society, for the sake of… not even making money, in some cases. Just numbers. We know a lot of these companies aren't making money—we're just collecting data. And at what cost?”

5.5.2.4 Labor pains: Exploitation, isolation, and burnout

Finally, participants reflected on challenges related to worker experiences, including exploitative labor arrangements, organizational siloing, and chronic burnout.

Overreliance on third parties and exploitative working arrangements. Despite limited resources, Trust & Safety workers and related teams are expected to support continual platform expansion and scaled growth, resulting in what participants described as an overreliance on third-party labor. Technology companies have increasingly outsourced content moderation and other data labeling tasks to third-party vendor workforces, often in regions with lower labor costs (Gray & Suri, 2019; Roberts, 2019). Despite strict production quotas and regular exposure to disturbing content, individual contractors earn relatively low wages, without the benefits and job security associated with full-time employment. “Yeah, people are getting paid… for doing really, really hard work,” said P2. “They’re not getting paid enough. They’re not being offered the benefits that they should be offered to do this very, very difficult work.” P3 agreed, noting the disparity in working conditions between content moderators and other technology workers:

> “For content moderators, value them like you do an engineer, in terms of compensation and benefits. Treat them as human equals, as opposed to disposable people that you expect to get burned out from seeing harmful content. Value mental health and resiliency for these people that are actually having to view the content every day.”

Participants expressed concern with the sustainability of outsourcing work that is critical to everyday business operations, in part because countries with lower labor costs are often located in regions more vulnerable to climate-related disasters. “Last year, there was a hurricane in the Philippines,” said P23. “There was no power at all, throughout the whole country. There were no reviewers who could do anything. That is a very real problem, and irresponsible to ignore.”

Companies may also expect that certain individuals will contribute their services on a voluntary basis. Civil society participants expressed particular frustration at the amount of work required to engage with platforms on behalf of users, which P22 described as “the free labor that people like me end up doing for tech companies.” P22 continued: “As Trusted Partners[8], we have to mediate and escalate things that we see on the platforms, or they won't be taken down—really flagrant issues, like death threats, rape threats, homophobic content in different languages. We are supposed to do the content moderation work and escalation work to our public policy or human rights contacts at these companies, or it won’t be taken down.” In addition to advocating for users and flagging specific content for potential removal, civil society groups may also provide more foundational governance labor, often without compensation. P22, who compiled dictionaries of potentially harmful terms using regional expertise, expressed doubts about whether or not this work was ultimately implemented:

> “As someone from the Middle East, I have had to build lexicons for homophobic and transphobic language. We had to bring in groups from different countries, to build these lexicons for Facebook—but then we never see them applied. They say, ‘Okay, we'll take this stuff and incorporate it in our language models.’ I understand it takes time, but we do this work for free. Again, I just need to reiterate: for free. Because they don't want to invest—or they aren't necessarily as serious about investing—in language models and context-driven policies related to hate speech. We end up doing all the work and becoming traumatized.”

Fractured understanding: Navigating organizational silos. Outsourcing critical work also creates extreme siloing, where fundamental functions related to content moderation and broader platform governance are separated from their closest partners. When policy teams are organizationally distant from their operational counterparts, it results in what P23 described as “the gap between policy and protocol,” which becomes particularly pronounced at scale. P23 continued:

> “Policy basically represents intent—what some platform considers harmful, potentially in consultation with some external experts. But then you need to translate that into a series

[8] Meta maintains a network of Trusted Partners, comprising “over 400 non-governmental organizations, humanitarian agencies, human rights defenders and researchers from 113 countries around the globe” (Meta Transparency Center, 2023).

> of steps for the scaled review team. You can't just tell them, ‘Hey, go nuts with your intuition,’ because there would be no consistency whatsoever. That series of steps can have large or small gaps with the intent of the policy.”

Participants also highlighted the impact of distancing full-time employees from their more precarious peers, including third-party content moderators and other contingent workers. “Contractors are partitioned off somewhere else, where the rest of the company doesn’t have good visibility and empathy on what they’re doing,” said P27. Industry participants described the daily experiences of content moderators as inaccessible to other employees—and, as a result, largely unaccounted for in the development of moderation policies, processes, and products, despite moderators possessing the most direct and timely knowledge of user dynamics and developing trends. “Content moderators don’t feel they have a voice,” said P22. “Even though they interact with this content the most and notice all the trends.” For non-staff content moderators—for example, those users who voluntarily moderate a subreddit or Facebook Group—this gap is even more significant. “Community admins don’t work with the Trust & Safety team at all,” P6 said of Reddit. “So there’s like, the people on the ground, working with the actual moderators, and then a whole other team working on what they call ‘Anti-Evil Operations,’ which is the system they use to implement site-wide community standards.”

For content moderation to truly succeed at scale, participants expressed a need for what P14 described as “a more holistic system.” P12 agreed, emphasizing the distance between user-facing product teams and the communities their products ultimately serve: “In the context of building a new tool, that is typically done by a team of engineers, led by a product manager—with very little input from the people who do the moderation work.” While organizational distance between relevant teams is one key barrier, systemic understanding is also restricted by the fragmentation of outsourced labor and the precarity of contingent employment. P12 continued:

> “Even when such input is solicited, there are other kinds of organizational barriers to getting good information—including the fact that when you do a site visit to a moderation outsource vendor, you get the feeling that employees don’t really feel free to express themselves fully, maybe under pressure from management. Of course, a bigger problem is that because of the atomized nature of the work they do, they might not even

> understand the types of questions that are being posed to them by the people who want to build tools, so there are several barriers."

Burnout, turnover, and institutional knowledge loss. Left with inadequate resources and facing relentless growth, individual workers—whose efforts to ensure user safety at scale present challenges under even ideal conditions—are left feeling unimportant, overburdened, and isolated from their closest peers, resulting in what participants describe as chronic burnout and frequent employee turnover. "There's a very high burnout and turnover rate for content moderators in general, but also for the people who work on these topics at the company," said P24. "At Facebook, a lot of people don't last more than a year on some of these teams, or leave the company altogether."

While the mental health and overall wellness of front-line moderators has garnered increasing attention in recent years (e.g., Roberts, 2019), participants expressed similar concerns for Trust & Safety workers and others in comparable roles. Although these workers are typically (though not always) full-time employees—and as such enjoy significantly more institutional support than their outsourced peers—they, too, conduct challenging work with insufficient resources. In addition to performing their role-specific responsibilities, workers may face challenges to their emotional and psychological well-being, such as what P6 described as "emotionally managing how you feel after dealing with difficult content." Participants indicated that Trust & Safety workers may also face unique administrative barriers, with some organizations requiring additional approval processes before certain employees can proceed with potentially sensitive work. "You're getting pushback from Legal on research or projects that you want to launch that could have huge impact," said P24.

These disproportionate impacts contribute to difficulties recruiting and retaining talent. "It does take time to build out teams—and to keep people on teams," said P24. "Burnout prevention for FTEs and content moderators… we know this is very difficult stuff to work on." P2 highlighted the difficulty of retaining expertise when individual workers regularly move between organizations:

> "It didn't matter how many people platforms were hiring, especially at the major companies. There just never seem to be enough people hired. I've heard more recently at the majors that there's a huge retention problem as well. This is a new field. These

> companies, as they're building up, they're trying to get talent—and so they're pulling from all these other companies. It's creating a pipeline problem. There just doesn't seem to be enough people, in terms of creating policy but also enforcing it. It doesn't matter how many tens of thousands of people are hired; it never seems to be enough."

In addition to problems retaining talent and appropriately resourcing critical teams, companies experiencing rapid growth may favor hiring processes that predominantly emphasize process efficiency, at the potential expense of identifying highly skilled candidates with domain-specific expertise. "When companies scale, they inevitably sacrifice on hiring requirements and training," said P31. "Part of the problem is, yes, it's a young industry. I worry over time, when things balloon, you have less focus on training—less focus on hiring people who can think through things in a nuanced manner."

Ultimately, participants emphasized that technology companies do not appropriately value work relating to user safety. Several participants drew parallels to other care-based industries, such as healthcare and domestic labor, where essential work is similarly undervalued. P16 expressed "wanting to see care work valued more highly; things like the 'Wages for Housework' campaign. Trying to push for the valuing of what would otherwise be under-resourced labor." P12 agreed, lamenting what she described as a general "undervaluing of the importance of the work—and workers." "Sometimes, some of the employees—and I think especially leadership—tend to forget that these are actual people doing the content moderation," said P24.

Executives enjoy disproportionate influence. While individual workers suffer the implications of increasingly precarious work, company executives enjoy disproportionate institutional power—including significant influence over individual content moderation decisions. Participants suggested that certain company leaders may overestimate the relevance of their own experiences and instincts, ultimately influencing organizational direction and broader strategy. "I remember very clearly when Dick[9] had gotten feedback from all of his friends that Block wasn't a feature that Twitter needed," said P9. "They really just needed Mute, so they should roll out Mute and get rid of Block at the same time. I was like, 'That is the dumbest idea,

[9] Dick Costolo is a former Twitter COO (2009–2010) and CEO (2010–2015).

and this will end in tears for many people.’ It only lasted like six hours after rollout, but it was very illustrative.”

When creating visual representations of current scaled content moderation processes, P8 and P9 highlighted the influence of certain “priority input,” including what they described as “random issues encountered by friends of the C-suite.” Other participants reflected on the influence of governments and other political actors on executive decision-making. P3 described ethical challenges faced by workers when asked to act in ways that conflict with their team’s stated mission and goals:

> “I think this is also something that is hard on ICs. I worked on escalations for a couple of years when I first started. A lot of times, we’d get escalations from foreign governments asking us to take down ‘fake accounts’ that were critical of the government. There were numerous examples of situations where leadership—I’ll be intentionally vague here—would say, like, ‘Do this, because we need to appease the government.’ And everyone who is an IC working on this was like, ‘What do we do? This goes against our team’s premise. This isn't a fake account.’ We're being asked to take it down as a fake account, and it's really only because it's an account that's critical of the government. And depending on the government, the company would push back or not push back. Morally, everyone always felt very conflicted about it.”

Participants emphasized a disconnect between the teams responsible for identifying risks to user safety and those with sufficient authority to make strategic decisions. “Ultimately, the communication flows one way—it comes top-down,” said P14. “It’s super hard for engineering or policy to convince CEOs of anything, based on any kind of bottom-up understanding of a problem.” In the absence of meaningful input, individual workers are often left to implement or even justify decisions they had no power to influence.

5.5.3 Content moderation futures: What can be done?

After identifying numerous ways in which contemporary technology companies fail to successfully govern their platforms, participants considered potential solutions.

5.5.3.1 Embrace holistic visions

In order to achieve sustainable success, participants encouraged technology companies to adopt a more holistic vision for the role of content moderation in broader community health. Participants highlighted that contemporary moderation systems are largely reactive, with the majority of governance-related resources allocated to the adjudication of individual content decisions—a phenomenon P8 and P9 described as "the 'whack-a-mole' element." P10 agreed:

> "Moderation systems are really just thinking about individual content decisions, and not like, what are the other things we're trying to promote? What does promoting health actually look like? I think that that is a big blocker, at least in the companies I've seen. They are pretty overwhelmed with having to make that decision, you know, hundreds of thousands of times, over and over and over again, in this kind of never-ending queue."

Instead of investing in reactive strategies and short-term fixes, participants emphasized the need for company leaders to clearly articulate a holistic vision of product safety and health. "There's just constant fracturing in the moderation space," said P33. "There's fracturing within a single area; there's fracturing between areas. There really does need to be a top-down vision that pulls it together. Leadership has to have a long-term, 'This is the thing we're trying to build' view—and it needs to be one that doesn't just rely on a single solution. Even though my colleagues have attempted to offer that repeatedly, I have yet to see someone in leadership adopt a vision like that at a social media company."

Moderation is essential. Participants believe the success of scaled content moderation relies on a paradigm shift: one that recognizes moderation as an integral component of the product experience, rather than a crisis response or otherwise necessary nuisance. Industry participants observed resistance to the characterization of content moderation infrastructure as foundational platform architecture. "Leadership is a little bit adversarial," said P33. "They view moderation as something that needs to be done—as a last resort, rather than an essential." P12 agreed: "I would reorient leadership to understand that good moderation is, in fact, a value add—it's not simply a burdensome 'cost center' to a company. It could be reconfigured and thought about in a better way—to actually be a selling point for a platform, rather than something to be hidden and swept under the rug."

Elevating the position of governance labor also requires improving the working conditions of individual employees. “If I could wave a magic wand and just change people’s mentality, I would do things like remove stigma and organizational barriers,” said P12. “The first thing I would probably eliminate is the system of outsourcing this labor in the first place, and pay it an appropriate wage for the skill level that it requires, which is very high.” P14 agreed, highlighting the ways in which low wages and limited career advancement further marginalize content moderation work at the organizational level:

> “The standing of content moderation within companies… it's an afterthought. It’s like, we have growth, we have engineers, we have sales, and then we have additional ‘also needed, but not important’-type things—like catering, housekeeping, and content moderation. Like, someone has to deal with the complaints, right? There isn't really a vision of the importance of community, or content moderation’s place within the actual product—or a vision of a social network at all, considering that these are the biggest platforms for humans to communicate. That doesn't seem to have resonated with any of the leaders of these organizations. And that's also seen in the career path of the content moderation guys. You have these super difficult and nuanced policies that are to be applied, ideally with 100% accuracy, by some outsourced people who get low wages. It's debatable if that's ever going to work.”

This structural devaluation is intensified by broader labor and equity concerns present across industries. “There are some macro-economic changes that need to be made,” said P2. “What we’re basically trying to do is figure out how we can create an equitable system, where maybe we have fewer of these problems. People are feeling a little safer, we have fewer divisions, so there are incentives for equitable and ethical moderation. Things like higher wages, the 3- or 4-day work week, paid leave… other ways of introducing equity in society. Universal healthcare. Broader socioeconomic changes would be helpful—not just reproducing inequitable conditions.”

Moderation is forever. Participants emphasized that content moderation, like other forms of governance, is a perpetual necessity—it can never be permanently solved. “There’s a sort of, ‘Oh, we have this problem solved, so we can stop worrying about it,’” P8 said. “Not realizing that it's an adversarial relationship. You never have that problem solved.” P33 agreed:

> “Leadership is thinking that eventually they can stop—like if we invest enough, we’ll have some magical solution, and then we can walk away from it. But moderation is forever fundamental to the product. It will have to be maintained forever. There’s this weird magical thinking about it that I wish would go away.”

Industry participants felt that company leaders largely desire ‘silver bullet’ solutions, even for complex sociotechnical problems that lack a simple, scalable fix. “There's a lot of like, ‘Everything is on fire,’” said P11. “Leadership wants a single solution—once you have it, it's done. It’s essential to move toward Trust & Safety people engaging with content moderation concerns as part of their product development. Moving toward things not just being on fire, but where employees are in a sustainable place; where they're actively and persistently building toward preventing harm, with organizational and financial support.” P33 agreed, advocating for more integrated approaches to the design and management of safety-related products and interventions:

> “There are logistical burdens that show up in the health space that don’t get addressed—like the fact that companies have more than one product surface they’re trying to solve for. They let the surfaces solve their problems separately, so there’s often a lot of noise and siloing. They’re designing one-off strategies—like ‘Hey, we’ll remove the content’—when no single strategy is going to be effective by itself. It needs to be a set of them, but people struggle to think beyond the one strategy on their mind. The way products are getting built puts a lot of additional burdens on the people actually attempting to build these systems. A lot of the leadership involved really fails to provide—or even allow the development of—holistic missions that would help tie together different parts of products and unify how they’re working.”

Moderation is communal. Participants described moderation as a critical mechanism for cultivating community, which requires consideration for a variety of individual experiences. “Content moderation has to be about the safety and the health of the greater community,” said P7. “You should care about every individual who is involved.” Promoting health at the community level requires consideration for the broader impacts of an individual’s actions, which may not always align with their intentions. “For successful moderation at scale, it doesn’t matter how the content conceivably could be understood," said P1. "What matters is how people are actually understanding it." P7 agreed:

> "You see people who are like, 'I didn't mean it that way.' I get it, but I cannot care what your intent was if your impact was that everybody had a bad day. That's a problem. If I'm moderating 1,000 active chatters, I cannot care if one person is having a bad day. I have to respect the fact that there are 999 other people who would like to have a good day."

Participants also discredited the notion that content removal is tantamount to censorship, emphasizing the role of moderation in maintaining an environment conducive to safe expression. "The inappropriate positioning of content moderation as an issue of 'free speech' is derailing and inaccurate," said P12. "People in positions of power conflate 'free speech' with a right to be abusive, violent, and toxic towards others," added P21. P33 compared content moderation to other necessary maintenance of public spaces:

> "Moderation promotes free speech. I brought up the public park metaphor: you want people to use your public park? Okay, well, people don't want to walk on the grass if there's glass in it—so you limit activities that will make glass in it. And that actually makes more people want to go there and do more things."

5.5.3.2 Design platforms that encourage prosocial behavior

Participants stressed the role of platform design in shaping user behavior, particularly for preventing or lessening harm. Because current approaches to scaled enforcement fall short of stated goals, participants described alternative approaches influenced by other governance contexts, such as criminal justice reform.

Establish clear expectations for appropriate behavior. Instead of focusing the majority of governance resources on policing rule-breakers, participants encouraged platforms to first establish clear expectations for appropriate behavior. "How do we shape community behaviors and norms?" asked P17. "There's a lot of focus on rule-breaking and the consequences of that. We have less on the norm-shaping side of it." While certain users may knowingly violate platform rules, some misbehavior can be prevented by ensuring that users understand normative expectations. P4 said:

> “Do research on what the rules should be, how they should be presented to people, and how people learn them. What would a moderation system look like if the biggest goal was to help people learn the rules? That is clearly not the intended goal of the majority of moderation systems right now, particularly at scale; the goal is for there to not be bad content on the platform. Helping people learn the rules could be a way to accomplish that, but it’s seen as a more roundabout way. The typical way to accomplish that is to identify and delete that content—and if people happen to learn the rules along the way, okay.”

Users are regularly exposed to normative signals that influence their understanding of a platform and shape their future usage, including their interactions with other users. In the absence of explicit guidance about platform rules and expectations, direct observations of others’ behavior will be the predominant normative influence. P10 emphasized the importance of onboarding experiences in setting clear expectations for new users, who are often hastily ushered into creating and engaging with content:

> “It starts when someone joins a platform or some new space. It's hard to talk about the stuff that happens at the very end—the moderation—without thinking about how they got there. You join a space, and you are shown some set of rules. Sometimes that's like, ‘Hey, go read our rules,’ or sometimes, ‘Here's our Terms of Service; click Accept,’ but rarely is it nuanced onboarding, where someone is guided through what the norms and expectations are. And pretty quickly thereafter, you're encouraged to create some content, whether that's filling out your profile or creating a post. Equally quickly, you're exposed to a bunch of content—‘Hey, here's a bunch of people you should be friends with or connect with.’ A lot of it is just observing what other people are doing, because people often don't jump in and just start creating a bunch of stuff. They're often just lurking, at least for the beginning, and starting to see how other people are engaging in this space.”

This brief exposure to platform rules also typically occurs before new users have enough knowledge of the platform to appropriately contextualize them. “The first thing that happens related to moderation is that, as soon as you join, you are—in theory—shown the rules,” said P4. “You don’t learn the rules. You’re shown the rules… maybe. It's so divorced from when you actually create content. It's not like when you post your first piece of content, then you're told about the rules.” Because new users also lack an established network and personalized feed, their first experiences of a platform will largely be characterized by algorithmic recommendations of content that is engaging, but not necessarily normatively appropriate or even typical. P10 said:

> “What if there was more curation around people's first exposure to the space? Most of the time, you're just thrown in—you're exposed to whatever you're exposed to. And on platforms that have algorithmically filtered feeds, you're being shown content that tends toward, like, salacious or controversial. The closer things get to pushing up against the rules, the more engagement and eyeballs it tends to get—so that tends to be what people are seeing. That doesn't seem to be conducive toward helping people learn what the rules are.”

Promote rehabilitation, not retribution. Because expectations for appropriate user behavior are often unclear, participants felt that platforms should prioritize user education over potential penalties. “I would love for content moderation systems to incorporate a ‘user education first’ approach,” said P25. “With the exception of high-severity harms or malicious intent, the system should not default to enforcement first, but to training or education.” P6 agreed, highlighting the lack of opportunities for offending users to understand and correct their behavior: “Part of what’s challenging is that there seem to be so few opportunities for people to learn from their mistakes. For people that do end up violating a policy or norm, it can be really challenging.”

Many contemporary platforms also lack specific incentives to encourage or reward desirable behavior, further contributing to uncertainty about normative expectations. “There’s a lack of incentives for users to behave well,” said P20. “Especially in commercial content moderation, which often uses punishment. On Reddit, they have awards and karma to incentivize users to behave well.” Given that users lack definitive guidance about appropriate conduct, participants felt that certain violators deserve more leniency. P25 advocated for a “second chances approach” for first-time violators, who rarely reoffend:

> “Many users who violate do so once—and after some education, do not continue to violate. This suggests that there should be the opportunity for users to ‘undo’ the

> problematic behavior, before any enforcement or strikes would kick in. If this hasn't occurred within, like, 24 hours, the platform steps in and a strike is applied."

P10, who recommended "more research about the experience of people who break rules," agreed that most policy violators are not repeat offenders—and discouraged platforms from ascribing malicious intentions to every user who runs afoul of their policies. "I'm constantly in these rooms where people are talking about 'bad actors,'" said P10. "I've done multiple recidivism analyses across different platforms that show it's an exceedingly small percent of people, despite how much room in the conversation they take up."

Favor flexible interventions. Not all rule violations are equal, and participants also suggested that platforms should implement more contextual interventions that better account for differences between violation types and offender motivations. "It has to do with context," said P18. "How do we understand a piece of content without also looking at the user, and what it might mean for that particular user in their own context?" P6 advocated for penalties that are proportionate to the specific offense, including the violation history of the offender:

> "Depending on what the violation is, context is taken into consideration. Different outcomes depend on what the violation actually was, and what kinds of information we know about the user. Somebody who's repeatedly violating a rule is probably not going to be dealt with the same way as somebody who's done it for the first time, or somebody who is brand new to the community. The severity of the infraction, the potential for harm that it causes… all of that is going to depend."

Uniform penalties also fail to account for differences between harmful behaviors, with acute behaviors often treated identically to those that progressively worsen. "We know that there are some problem areas—like eating disorders, suicide, self-injury, radicalization and others—that can have an escalation trajectory," said P25. "There is a percentage of the population where their behavior will become increasingly worse over time, but the response by platforms tends to be the same each time there is a harmful event." P8 agreed: "Harm is usually a process, not an event—and we only look at single events."

Participants also felt that many platforms rely on overly blunt interventions, which P9 described as "a historical over-indexing on content removal as being the sole type of content moderation." "Moderation is often focused on very binary output," agreed P30. "Leave up and remove." By focusing on binary removal decisions, platforms may lose sight of broader governance goals, such as promoting civil discourse or reducing user harm. P10 said:

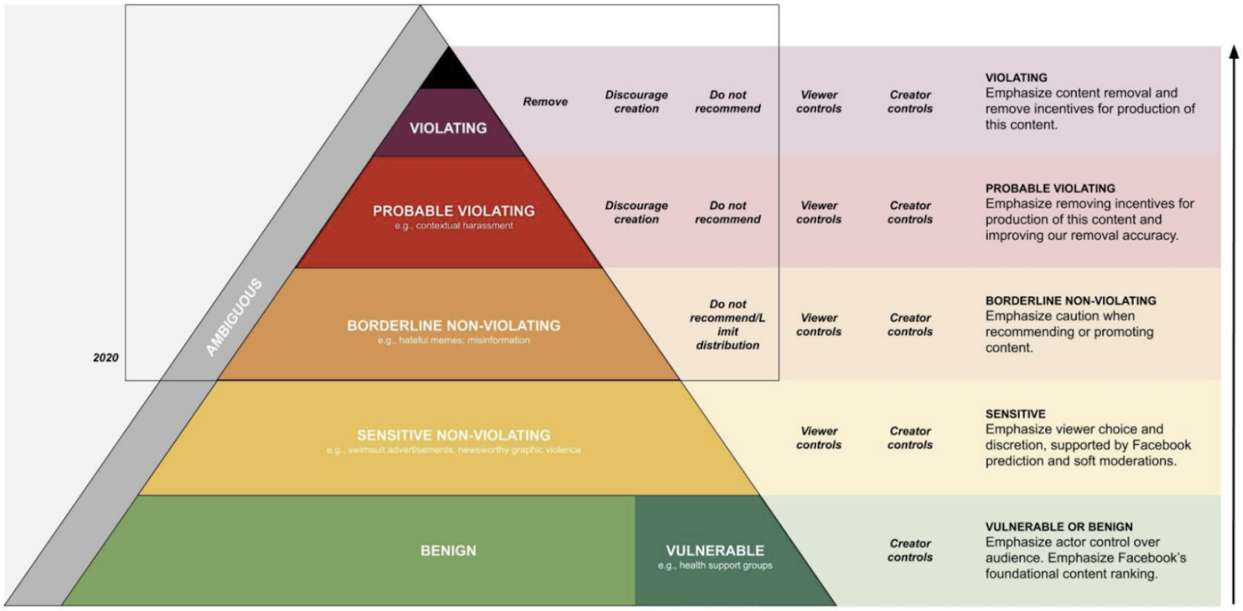
> "There is an orientation toward this binary of moderation, where stuff is either good or bad. Especially in the actual operations of content moderation, binary decisions are being made: stuff stays up, or it gets taken down. I think it points us toward the wrong goals. There are many other goals a moderation system can work toward. Trying to address harm is something that doesn't really serve, when you're just thinking about whether something stays up or goes down."

Ultimately, participants stressed that effective interventions encourage offenders to reform their behavior. "Effective sanctions actually correct behavior that is untoward," said P27. "We struggled at Facebook coming up with ideas that actually move the needle toward creating a healthy community and aren't just overly simplistic or heavy-handed… that end up doing more damage than good."

Implement responsive regulation. P33 shared a pyramid model influenced by Braithwaite's (2002) responsive regulation (see Figure 5.2), which allows for the deliberate application of adaptable sanctions with escalating penalties to encourage cooperation. "This is another way of thinking about different types of content—a holistic vision," said P33. "Not just using a single strategy to moderate content, but using a collection of strategies, and gradually layering on more strategies as risk of harm becomes more severe. It's very focused on a more restorative justice view." Because responsive enforcement systems provide offenders with opportunities to adjust their future behavior, users who repeatedly violate rules are clearly signaling their intention—creating justification for the application of harsher penalties, such as device-level bans. Instead, most contemporary platforms favor blunt enforcement systems, applying uniform penalties to every offender.

Participants noted that uniform penalties are both too harsh to effectively reform well-intentioned offenders and too lenient to deter chronic recidivists. "There's a lack of ability in some cases to tell whether you're dealing with somebody who really just needs rehabilitation and is sort of… fixable, let's say," said P29. "Or if you're dealing with somebody who is there to game your systems, and it's like the seventh account that they've created; they're just maliciously wasting your time and engaging in bad faith. When you have 'one size fits all' systems, you end up struggling to design systems in a way that optimizes for both use cases—and you end up pleasing neither." P32, who encouraged platforms to "crack down harder on repeat offenders," suggested that platforms who fail to appropriately penalize frequent violators may facilitate their continued influence:

> "We've been working on this project about what happened to the election lie superspreaders—who were the biggest spreaders of election fraud in 2020, and what ever happened to them online? What we learned is only a third of them were banned. The rest have gone on to be completely divisive in a series of issues that have divided the country since. They went on to be influential in conversations about grooming and critical race theory—and yet, there was an opportunity where they really did break the rules in a grave way. If there were a harder crackdown on those particular megaphones, there would be a different conversation."



VIOLATING
Emphasize content removal and remove incentives for production of this content.

PROBABLE VIOLATING
Emphasize removing incentives for production of this content and improving our removal accuracy.

BORDERLINE NON-VIOLATING
Emphasize caution when recommending or promoting content.

SENSITIVE
Emphasize viewer choice and discretion, supported by Facebook prediction and soft moderations.

VULNERABLE OR BENIGN
Emphasize actor control over audience. Emphasize Facebook's foundational content ranking.

Like other models of responsive regulation, we increase distribution friction as we move up the pyramid.

Figure 5.2: An enforcement model influenced by Braithwaite's (2002) responsive regulation.

5.5.3.3 Incentivize corporate accountability

Existing corporate structures have failed to adequately protect users' safety. Participants stressed the need for greater corporate accountability, whether incentivized via regulatory structures or alternative levers. "Companies are not accountable," said P20. P32 agreed:

> "I do think there is a bigger role for governments to play. Private companies are in the business of making decisions they probably shouldn't be making on their own. Too much of the public square is being outsourced to these private companies. It's become a job—but these are major societal issues that historically governments would play in much more."

Regulation: Promising, but not a panacea. Participants described regulation as a critical mechanism for encouraging corporate responsibility, including basic transparency into company practices. "When I first started reporting on this in 2017, Facebook wouldn't even tell us that they had content moderators," said P32. "It was so unglamorous, it was like they wouldn't even acknowledge it—until the Congressional hearings." P31 noted the effectiveness of financial penalties in motivating compliance with child safety laws and other public interest policies:

> "Capitalism could potentially remove incentives for tech companies to invest in certain areas. If you've seen Silicon Valley, there's a famous episode where one of the characters almost bankrupts the company by messing up COPPA, the Children's Online Privacy Protection Act. Basically, there are massive fines from the government if you are found to not be deleting people under the age of 13 from the platform, or if you are negligently allowing them onto your platform in a way where you're collecting their data. I would love for there to be more laws that motivate companies to remove more misinformation, to be more proactive… to maybe even discourage companies from existing, if they cannot exist in such a way where they're not causing societal harm. I'm in favor of government intervention, while being mindful of the risks associated with its misuse."

While participants expressed support for the development of new regulations to force more responsible corporate action, they questioned lawmakers' proficiency in a problem space rife with complexity and nuance. "Emerging regulation is becoming really important at the moment," said P25. "Not all of the specific regulations are actually fully informed on the realities and challenges of content moderation." This disconnect between policymakers and technologists may result in clumsy or ineffective regulations that fail to address the root cause of these problems—while simultaneously creating additional stress for resource-sparse teams. P9 referenced Australia's Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019, which passed through both houses of parliament in less than 48 hours (Douek, 2020):

> "I can tell you, from massive amounts of personal experience, the challenges involved in getting even a group of governments in a set of countries that are generally aligned to agree on anything within a time span of less than, I don't know, a two-year period. It's like herding cats on methamphetamines. And the cats are going to also just leave and then do something like what Australia did—they wrote a law and then passed it overnight, basically, in the aftermath of the Christchurch shooting."

Participants also expressed discomfort with the potential for government overreach and corruption, particularly given the surveillance value of user data. "Do we want to be outsourcing this to governments?" asked P32. "Governments around the world have deep interests in targeting citizens for certain reasons. Many governments shouldn't be trusted with that." P9 replied:

> "I think that's a robust understatement. If you look at the rankings for press freedoms, even the countries where theoretically they're quite high, you still have almost inevitably bad actors within those systems who could still be trying to push for removal of content that is completely legitimate—whistleblowers, activists, dissidents. We've had examples of governments straight up telling us untruths to try to get content removed, like claiming there wasn't a protest taking place. There absolutely were protests happening. The government was fully aware of them, and they were just trying to shut that down before it got out further."

P31 agreed, acknowledging the political and moral complexity of government intervention, which may effectively reduce certain harms while also creating additional risks:

> "It's very complicated. Should we potentially have laws or even higher-level governmental systems in place that restrict access to certain content? I hate that idea. I read 1984. But at the same time, I look at countries that do have some restrictions in place, and their CSAM risk is almost eliminated in some cases—so I do think that some government intervention might be necessary. I worry that tech companies are sacrificing

> the well-being of society for greed; for making money. I want more government intervention and more laws, without those laws or tools being exploited by corrupt government officials. I don't know what the solution is."

While external incentives may be necessary to reorient broader corporate priorities, internal change is still needed. Regulation is a powerful lever for motivating company leaders to dedicate resources to areas of their businesses they may otherwise neglect—but as P7 noted, systemic transformation requires cultivating additional understanding and commitment:

> "We talked about finding ways to incentivize companies to prioritize users—and user safety—over corporate profits. Some of this can be government pushes; laws and guidelines, like 'If you don't do this, we will take your profits' or whatever. If you incentivize companies to do that, you can then take the time to do things like educate—everybody at a company, not just moderators, but all the way up to your C-suite. Why is content moderation important? Why is nuance important? You can have incentives to make companies not care as much about their profits, which is great, but if there is no training—if there's no push for nuance—then you lose any kind of forward momentum that you get from having laws that will inflict punishments on folks."

Participants also cautioned that technology companies may be incentivized to identify loopholes to avoid increased scrutiny by governments and other regulatory agencies. "I truly believe that Facebook chose to go to end-to-end encryption to avoid regulation," said P21. "It had very little to do with protecting people who are vulnerable; there are plenty of other tools that journalists, activists, and other people in marginalized situations can use. They wanted end-to-end encryption for Messenger to avoid responsibility."

Leverage alternative accountability mechanisms. Creating true accountability for modern technology companies will require leveraging a variety of mechanisms. In addition to pursuing regulatory solutions, participants emphasized the importance of public scrutiny, including negative media attention. "The incentive for firms to do good content moderation is really a soft power incentive," said P12. "Sure, there are government regulations—and there are more and more coming online—but outside of the cases that those regulations address, which are a fairly small percentage, what really ends up being the incentive for firms is things like bad PR, or a risk of losing their userbase.” P2 agreed, noting that profit can be an effective lever for motivating corporate responsibility:

> “Commercialization adds some good things towards this dynamic. Some of the ways we’ve really gotten movement on a lot of moderation issues is because of like, trying to keep advertisers happy. That can be both bad and good; it just depends on which groups you're making happy, and those commercial logics. The real incentive is to keep people on the platform.”

Pressure from users can be an effective lever, too. “Reddit describes it as a process of maturing,” P2 continued. “Actually trying to figure out what these dynamics should be, and drawing some clear lines around subreddits like r/fatpeoplehate or r/beatingwomen. Really what that was was a lot of pressure from the community. Pressure from advertisers and things like that.” Still, participants recognized the stark power differentials between even organized communities of users and other corporate stakeholders, such as investors. “There’s nothing really binding platforms to be responsive to the communities that get harmed by them,” said P5. “Nothing like the way they have to be responsive to their investors and SEC filings. There’s nothing that’s really tightly coupling those groups.” P12 agreed:

> “In terms of users, the leverage is just not there in an organized way—in the way it is there for shareholders. The case study of this power imbalance is what has gone on with Twitter, because nothing about that was good for users. Nothing about that was good for the public. And yet fiduciary duty to shareholders was what guided that entire process, and what will continue to guide it. We see a real misalignment there.”

Ultimately, corporations wield power that individual users lack, and participants emphasized the importance of structured coordination and collective action in provoking meaningful change. Persuading technology companies to be more accountable to their users may require creative solutions for consolidating individual influence, such as what P5 described as an “eyeball strike”:

> “Imagine all the users of Facebook somehow unionize and drive revenue to zero. I was thinking of ways to upend the system. If you have no accountability to the communities that you harm, unions are a historical example of an approach to mitigate that.”

5.5.3.4 Explore alternative models

Corporate accountability is not guaranteed, and participants also emphasized the need to explore alternative technology models that allow for greater transparency, experimentation, and user empowerment.

Forsake the 'killer app.' Participants reflected on the fundamental tension between establishing consistent, principled governance and the pursuit of continual expansion and growth. "Platforms are trying to operate so broadly, in terms of who they want on there," said P2. "It makes it very difficult to create one rule. You could be alienating lots of different parts of your community."

When technology companies design platforms intended to serve heterogeneous populations of global users, certain values are necessarily prioritized over others, resulting in governance practices that inevitably marginalize some users. "Many of these platforms are designed to create one large public space," said P17. "And the consequence is that there are certain groups that are more privileged than others. It creates lots of weird disincentives for effective moderation." Rather than committing to specific principles, most platforms pursue impartiality, in the hopes of appealing to the broadest possible userbase. When they fail to establish universally applicable rules, they also compromise their ability to deliver consistent scaled outcomes. "There was a reluctance from platforms, understandably, to intervene in geopolitics," said P2. "So there was a problem with creating one kind of rule that could be applied more generally—which is a problem if you're trying to moderate at scale." P32 added:

> "I feel like it could be more efficient if each platform just said its values. That's actually one of the interesting things about some of the platforms that have emerged on the right. They say they're 'free speech' platforms, but in a way, what they are is a particular ideology—and they're okay with it. I know that would sacrifice users, if you didn't at least try to appeal to everybody… but there is no societal common denominator, and so admitting that, and then going 'Okay, well what's *our* common denominator, even though it's going to make a lot of people disagree with us.' Getting away from that approach of trying to be all things to all people would be helpful."

Participants also highlighted the difficulty of governing broad, general-purpose environments designed to serve a range of user needs. "We really run into issues with spaces designed for the flexibility to become many things," said P26. "Because we've instilled that flexibility, we need to collapse and standardize. In spaces that are designed for very particular outcomes, we can do a lot more to empower individual users or small groups to manage their own safety and to escalate requests for help—because those spaces don't suddenly change." Instead, participants encouraged companies to design technologies for specific purposes and audiences. "One thing that I would love to see a lot more of—and I think empowers a lot of better solutions for addressing disruptive or harmful content—is having platforms and spaces designed so that they are better calibrated for the needs of that group," said P26. "Maybe the fundamental problem is just that we're using the wrong platforms," said P29. "Maybe the mistake we're making is trying to find a way to make one platform be everything for everyone."

Empower community governance. Given the challenges associated with uniform governance of general-purpose platforms, participants advocated for more localized and participatory alternatives, empowering users to shape and steward their own community spaces. "Top-down content moderation interventions can cause harm," said P11, recommending "more empowerment on the user side." Even companies that utilize scaled, platform-level governance rely heavily on volunteer moderation labor, such as users who voluntarily moderate Facebook groups or subreddits. Participants recognized the unique burdens on these users, who spend significant amounts of their personal time interpreting user reports, coordinating with other moderators, and communicating outcomes to offenders—often without the benefit of automated enforcement tools or other operational mechanisms common in scaled moderation. P19, who researches user-governed Discord servers, encouraged platforms to implement moderator rotation programs, both to reduce the possibility for moderator burnout and create additional opportunities for users to participate in community governance:

> "People want to create their own community management—but they do seem to get trapped sometimes. They can't give up the power; and then, of course, a lot of people get burnt out. Incentives for rotating people in and out helps with a lot of the problems of moderating, especially at scale, for these kinds of self-governing spaces. And not just to solve the problem of delinquent moderators or burnout, but also… what would it mean for the community if everyone has an opportunity to help out and make it a better place?"

Participants reflected on the value of shared responsibility, noting the importance of community engagement in promoting ownership and accountability among individual members—in turn reducing potential misbehavior. “If I can just go to a community garden and pick up trash, that’s one way I could help contribute,” said P19. “I’m thinking of online spaces in a similar way. It doesn’t require monetary incentives. There is a stronger reward, in that I also get to have stronger participation.” Other participants questioned who ultimately benefits from this vision of collective community stewardship, particularly under platform capitalism, where user-generated content is heavily monetized. “I really love that,” said P2. “Except it would be so weird if Monsanto came into that community garden and started selling all of the vegetables the community was producing. That’s kind of the dynamic right now.”

5.6 Discussion

These results reveal that contemporary social media governance is broken—yet few companies seem to meaningfully invest in its repair. I argue that successful governance is undermined by the pursuit of technological novelty and rapid growth, resulting in platforms that necessarily prioritize innovation and expansion over public trust and safety. To counter this dynamic, I revisit the computational history of care work, to motivate present-day solidarity amongst platform governance workers and inspire systemic change.

5.6.1 Move slow and fix things

Modern technological development is shaped by the promise of “innovation,” informed by the techno-optimistic notion of technological progress as both inevitable and desirable (Avle et al., 2020). Innovation is tightly coupled with entrepreneurialism, which encourages individuals to contribute to broader economic development by promoting a democratizing ethos of invention, collapsing “vast gaps in money, formal knowledge, and authority” (Irani, 2019) to disguise the power dynamics that determine who will succeed and who fails (Becerra & Thomas, 2023). Innovation promises novel solutions to human problems, but also the production of wealth for individual innovators, complicating distinctions between public good and private gain. Participants in this study observed how this ideology manifests within large social media companies, describing corporate cultures that prioritize new product development over the resolution of existing product harms—reflecting an incentive structure that rewards speed, novelty, and growth over user safety.

Innovation is frequently framed as a self-evident social good, emphasizing disruption for the sake of progress and disguising the political and economic interests embedded in what, how, and for whom innovation occurs (Irani, 2023). By casting technological advancement as universally beneficial, the potential harms produced by unbridled innovation—such as labor displacement, algorithmic discrimination, and expanded surveillance—are obscured. As a result, the prevailing ideology of innovation normalizes risk-taking while externalizing responsibility for risk mitigation, reinforcing a model of progress that privileges economic ambition over public accountability and societal health. Participants described how this logic legitimizes underinvestment in platform governance, resulting in Trust & Safety teams who lack sufficient resources or institutional authority to produce effective moderation. Instead of recognizing governance as a platform's core function (Gillespie, 2018), key responsibilities are delegated to underpaid commercial moderators, unsupported community volunteers, and uncompensated third-party experts—practices that reflect a broader strategy of outsourced accountability.

In order to industrialize inventive production, innovation encourages frequent experimentation, resulting in the rapid deployment of technology products with minimal testing and limited attention to potential downstream consequences (Pfotenhauer et al., 2021). Commercial content moderation systems rely heavily on machine learning models, which participants noted are frequently released without robust internal understanding of their actual behavior—which may differ from their intended function. Even those responsible for the development of automated tools are unable to explain how specific governance decisions are made by the systems they built and maintain. This orientation toward efficiency comes at significant cost: inaccurate moderation, inconsistent policy enforcement, and the marginalization of users whose contexts and experiences fall outside the narrow training data used to calibrate these automated systems.

Participants similarly emphasized a lack of corporate investment in foundational governance infrastructure. While "new and flashy" projects receive attention and accolades from company leaders, participants noted that "hard but essential work"—such as maintaining or improving content moderation tools—receives little recognition during evaluations of individual and team performance. This orientation toward innovation disincentivizes investment in content moderation, system maintenance, and other essential safety infrastructure, areas characterized by company leaders as cost centers rather than engines of growth. The result is a corporate culture that rewards the continual creation of new features, products, and platforms, producing "innovation" at a pace that restricts even internal scrutiny—and resulting in the public release of experimental technology products with unknown ethical implications.

Success in innovation is commonly defined by how quickly and widely a product can be deployed, reflecting an underlying alignment with market-driven values. By prioritizing speed and scalability, innovation favors actors with existing resources and infrastructure, constraining which problems are solved and for whose benefit (Irani, 2019; Irani, 2023). Problems that cannot be solved through scalable solutions—or whose solutions primarily produce social but not monetary benefit—are deemed less relevant or even antithetical to corporate goals. Participants noted the "adversarial" positioning of efforts to mitigate risks to user safety, due to perceived delays in product development timelines or negative impacts to key performance metrics. More holistic, forward-looking governance strategies and pro-social platform designs—"innovations" rooted in social justice and collective success—are repeatedly sidelined, despite both theoretical promise and empirical validation (Tyler et al., 2021; Katsaros et al., 2022). Taken together, these findings reveal how the prevailing ideology of innovation externalizes the labor of governance and delegitimizes repair, rendering the systemic work of harm reduction peripheral to the business of building scaled platforms.

5.6.2 Countering the politics of scaling

Scalable platform technologies embody a broader "scalability zeitgeist," or the modernist preoccupation with technological solutions that can be efficiently replicated at exponential scales (Pfotenhauer et al., 2021). Contemporary ambitions of scale are articulated through entrepreneurial strategies such as "growth hacking," or the rapid acquisition of users to accelerate revenue growth—privileging speed and market dominance over product quality and longer-term sustainability. Economies of scale have long been a cornerstone of industrialization, exemplified by the mass production of standardized goods, enabling high output at low cost. The rapid industrialization of commercial content moderation is no different, and as platforms expand, they rely on scalable, standardized workflows and low-cost labor to process vast volumes of user-generated content quickly and efficiently, often at the expense of equitable governance (Gillespie, 2018; Roberts, 2019).

While scalability is often marketed as a sustainability strategy—with the purported aim of allowing businesses to handle increasing demand without incurring proportionally higher costs—

it signifies a particular vision of the future grounded in perpetual growth (Hardy, 2019), made possible by the externalization of social and environmental costs. In what Pfotenhauer et al. (2021) describe as the “Uberization of everything,” countless technology start-ups now aspire to disrupt existing markets by building scalable platforms that shift costs, risks, and other traditional business responsibilities onto others—namely users, workers, and governments. As social media platforms similarly prioritize growth and scalability, participants described chronic underinvestment in Trust & Safety teams and related functions, increasingly shifting the burden of platform governance to outsourced workforces, algorithms, regulators, and users themselves.

Scalability creates structural constraints on what kinds of governance are possible. Due to the demands of global scale, content moderation systems become reactive by default, producing an unrelenting stream of individual decisions. This Sisyphean effort—which participants likened to an endless game of “whack-a-mole”—requires extensive resourcing, impeding investment in more preventative solutions. These politics of scaling profoundly shape the governance capacity of contemporary social media platforms, as evidenced by the findings presented in this paper, which underscore the structural tension between global scalability and the lack of universally applicable standards for acceptable behavior. While effective governance is responsive to context—for example, local variation in norms, values, and interpretations of harm—participants describe ways in which scale inherently obscures difference, resulting in content moderation policies and practices that cannot adequately address nuanced harms. Instead, social media platforms implement crude solutions that predominantly address issues with broad consensus, in what one participant described as a “lowest common denominator” approach to moderation—producing governance outcomes experienced by users as arbitrary, inconsistent, or unjust.

Under what Srnicek (2017) describes as platform capitalism, platform technologies generate profit by building and controlling digital infrastructures that mediate interactions between users—for example, connecting riders with nearby drivers—while continuously extracting, analyzing, and monetizing the data those users produce (Hardy, 2019). By positioning themselves as intermediaries, platforms avoid the legal and regulatory responsibilities of direct service providers (Gillespie, 2010) while accumulating massive amounts of valuable user data (Zuboff, 2018), which they use to optimize engagement, predict behavior, and deliver targeted advertising in increasingly automated ways. These logics of extraction result in platform designs that tolerate or even encourage harmful content or behavior in pursuit of profitability, which participants described as a fundamental misalignment between business incentives and user safety—exacerbated by a culture of quantitative obsession that cannot meaningfully capture or effectively respond to complex, contextual harms.

Though the imperative to scale is frequently presented as a neutral or purely technical goal, this framing obscures the social, economic, and political stakes of widespread expansion (Hanna & Park, 2020; Pfotenhauer et al., 2021). Scalable governance demands uniform rules, and participants emphasized the disproportionate influence of Western perspectives on content moderation policies and practices, reflecting computing's intrinsic "colonial impulse" (Dourish & Mainwaring, 2012) and further marginalizing vulnerable users by reproducing global power asymmetries and entrenching structural harms. Through the lens of scalability, social media platforms become sites of contested governance, where decisions about speech, visibility, and harm are shaped by capitalist logics of extraction and expansion. Without a fundamental reorientation toward equity, accountability, and care, the structures that enable corporate profit will continue to undermine the conditions necessary for just and effective governance.

5.6.3 Reclaiming computational logics of care

As the present study demonstrates, improving the future state of scaled content moderation will require fundamentally reorienting how governance is framed—not as a reactive system of control, but as a continuous, contextual, and communal practice of care. Though dominant narratives of technological progress often invoke masculine, capitalist logics of innovation, disruption, and expansion (Irani, 2015a; Irani, 2019), technology is created and sustained by feminized and therefore politically devalued forms of care labor, such as support, maintenance, and repair (Fisher & Tronto, 1990; Jackson, 2014). Indeed, computing itself was built by care labor: NASA's core memory was constructed by "little old ladies" who hand-threaded thousands of wires through small magnetic cores, delicate craftwork requiring precision and patience—yet rendered "unworthy of remembrance" by the male engineers and astronauts who both depended on and were credited with its success (Rosner et al., 2018).

As Rosner et al. (2018) argue, the characterization of this highly skilled, error-intolerant work as manual ("feminine, menial, low-status") rather than cognitive ("masculine, innovative, high-status") labor demonstrates a lack of sympathetic context, or the situated knowledge required to appropriately value sociopolitically invisible labor (Star & Strauss, 1999). In the present work, participant experiences reveal that corporate technology leaders similarly fail to recognize the complexity of governance work, consistently undervaluing the labor required to responsibly govern online platforms. While participants characterized content moderation as an ongoing obligation, they questioned “magical thinking” from company executives, who expect scalable, permanent solutions—and who demonstrate a general reluctance to invest in any governance labor. In what one participant described as an “Oh Shit” model of governance, companies typically only establish a content moderation program following a crisis, instead of designing platforms with prevention and sustainability in mind.

Governance labor is also rendered structurally invisible by resisting the forms of measurement that typically justify institutional investment—and when value cannot be easily counted, it is systematically discounted. Because Trust & Safety work is ill-suited to quantification, corporate practices that prioritize the optimization of standardized metrics—used to communicate quarterly performance to company shareholders, but necessarily abstracting “the process of work being done” (Star & Strauss, 1999)—result in what participants described as chronic underinvestment in essential safety infrastructure. Despite content moderation tools that participants emphasized lack necessary context, the maintenance and improvement of existing systems is not prioritized, producing moderation outcomes that are at best inaccurate, and at worst, inequitable. Initiatives that do not readily translate into measurable business outcomes—such as increased user engagement or advertising revenue—are similarly deprioritized, reflecting a narrow conception of value that ignores the long-term benefits of community-driven design. These results reveal how foundational caretaking labor is rendered legible only as cost, not value, reflecting a broader political economy that separates socially reproductive labor—the labor of maintaining people, communities, and institutions—from labor considered economically “productive” under capitalism (Fraser, 2016).

Under financialized capitalism, short-term financial gain is privileged over long-term production, including social well-being—despite capitalist production relying on broader social capacities (Fraser, 2016). Social platform technologies exhibit a similar contradiction: in place of the slow, situated, and relational labor required to sustain healthy individuals and communities, technology companies privilege binary enforcement mechanisms that can be automated and quantified, mistaking efficiency and throughput for attentiveness and care. Content moderation professionals described current scaled moderation systems as overly reactive and reliant on blunt interventions (e.g., “leave up or remove”), despite workers’ intimate knowledge of—and desire to produce—more intentionally prosocial designs. As Seering et al. (2022) argue, content removal is just one element of "a deeper social process of nurturing, overseeing, intervening, fighting, managing, governing, enduring, and stewarding communities," a perspective reflected in the range of care-based interventions—user education, rehabilitation, responsive regulation, and so on—recommended by participants. Rather than merely policing infractions to achieve minimal compliance with regulatory requirements, participants expect technology companies to take seriously the work of cultivating, repairing, and sustaining the communities their products purport to serve.

This refusal to appropriately value care work is compounded by structural power imbalances, including labor precarity. Much like the women who built NASA's core memory systems, commercial content moderators are viewed as expendable, even as they absorb the psychological and operational burdens of the platform's most difficult problems (Roberts, 2016; Roberts, 2019). Participants recounted the precarious conditions of outsourced workers, who earn low wages, lack mental health support, and work remotely from regions with increasingly volatile climates. While contractors are physically and socially distanced from the designers and engineers who build the platforms they maintain, civil society partners voluntarily contribute expertise they rarely see integrated into public-facing products. Even salaried Trust & Safety workers conduct challenging work while organizationally isolated from their closest partners, resulting in what participants characterized as chronic burnout and frequent employee turnover. Platform governance workers operate in silos, with little influence over the systems they are charged with protecting—a structural fragmentation that reinforces the invisibility and undervaluation of care labor.

These findings reveal a platform governance ecosystem that depends on care labor but refuses to honor it, with complex, skillful, and emotionally demanding work obscured behind dashboards, outsourced across borders, or mistaken for a compliance task rather than a core function. Content moderation, like all forms of governance, is a manifestation of care, and it deserves to be valued accordingly. Reclaiming computational logics of care requires foregrounding the experience of workers (Irani, 2015b; Roberts, 2019), investing in maintenance and repair (Jackson, 2014; Schoenebeck & Blackwell, 2021), and building systems that reflect the relational nature of governance infrastructure (Star & Ruhleder, 1996; Gillespie, 2018). As feminist HCI scholars have long argued (Bardzell, 2010; Irani, 2015b; Dombrowski et al., 2016;

Rosner et al., 2018), to care is to attend to complexity, and to commit to the critical evaluation of technologies out of an interest in their material improvement—or as Puig de La Bellacasa (2011) writes, “we must take care of things in order to remain responsible for their becomings.”

One potential antidote to this ongoing crisis of care (Fraser, 2016) is the organization of social media governance labor, resisting institutional atomization and centering the workplace as an essential location for contesting the possibilities of governance. Aligned with recent calls for the promotion of a worker-centered HCI (Irani, 2015b; Dombrowski et al., 2017; Rosner et al., 2018; Roberts, 2019; Fox et al, 2020), these results underscore the importance of scholarly attention to the governance workers who build, maintain, and protect the platforms that structure our daily lives—and the need for individual workers to organize their collective power. Organized care labor is visible care labor, a critical first step in transforming scaled content moderation from a series of fragmented tasks into a shared practice with a common voice, collective memory, and strategic leverage. As HCI expands its ethical commitments beyond usability and inclusion, centering the working conditions and organizing capacity of governance workers across the “sociotechnical stack” (Qiwei et al., 2024) offers a pathway toward redistributing power and normalizing the value of care. Ultimately, while improving the future state of scaled content moderation will require a host of systemic changes, worker solidarity is foundational to achieving equitable governance and broader social justice.

5.7 Conclusion

This study broadens current understandings of social media governance by examining the lived experiences of practitioners who enact, examine, and engage with scaled content moderation systems. Through a series of participatory design workshops with content moderation professionals—from part-time content moderators and university researchers to corporate vice presidents—I produce a more intimate understanding of the complex landscape of people, practices, and politics that ultimately determines how contemporary social media platforms are governed. By characterizing platform governance as essential care labor—highly skilled, constant, and undervalued—this research highlights the need for worker solidarity and broader structural change to advance more equitable technology futures.

5.7.1 Acknowledgements

Thank you to Umang Bhojani for his contributions to data collection and workshop design. I am deeply grateful to my research participants for their competence, candor, and courage. I also wish to acknowledge those who felt unable to safely participate, and whose absence is not overlooked.

Chapter 6 Discussion

This dissertation broadens current understandings of social media governance through three empirical research studies, each examining the perception, classification, and governance of harmful behaviors in online spaces. Together, these studies provide a layered account of contemporary social media governance, demonstrating personal experiences, social constructions, and institutional regulations of online harm. Each study concludes by offering empirically informed and theoretically grounded recommendations for improving the design and governance of modern social media platforms.

The first study (Blackwell et al., 2017) investigates the sociotechnical implications of classification systems in the context of online abuse, drawing on interviews with 18 users of HeartMob, a private online community designed by and for targets of online harassment. The findings demonstrate that platform policies and support tools can both validate and invalidate individuals' experiences of harassment, ultimately influencing which experiences are deemed worthy of support. Echoing Bowker and Star (2000), this study emphasizes that classification is never a neutral or purely technical act, and classification systems that encode dominant values risk further marginalizing vulnerable users by reinforcing existing structural inequities. Informed by intersectional feminist theory, the study concludes by advocating for more participatory approaches to platform design, centering the perspectives of vulnerable users to produce moderation systems that more appropriately account for diverse experiences of harm.

When social media companies fail to effectively govern their platforms, users may pursue their own forms of justice. The second study (Blackwell et al., 2018) investigates *retributive harassment*, or the use of online harassment as a controversial form of social sanctioning. Through two online experiments (n=160; n=432), this study demonstrates that individuals believe online harassment is more deserved and more justified—but not more appropriate—when the target has committed some offense. This effect was stronger amongst individuals with a propensity for retributive justice, or the belief that offenders deserve sanctions that are proportional to the severity of their crimes (commonly referred to as “an eye for an eye”). Promisingly, this perception was reduced by exposure to a single dissenting voice—i.e., a bystander intervention—even amidst other conforming responses. The study concludes by discussing alternative approaches for responding to online harassment, including platform designs that promote restoration over retribution and encourage bystanders to safely intervene.

The third and final study examines social media governance through a worker-centric lens, drawing on six participatory design workshops with 33 content moderation professionals to illuminate the underlying logics that construct and sustain contemporary platform governance. While content moderation professionals broadly agree that successful moderation is principled, consistent, contextual, proactive, transparent, and accountable, they also describe numerous structural barriers to achieving these goals at scale, producing platform governance practices that are inconsistent, ineffective, and even harmful. I argue that successful platform governance is undermined by the pursuit of economic innovation and rapid growth, resulting in platforms that necessarily prioritize novelty and expansion over public accountability and trust. To counter this dynamic, I revisit the computational history of care work, to motivate present-day solidarity amongst platform governance workers and inspire systemic change.

Taken together, these three studies contribute a vision for social media governance inspired by theories and principles of justice, or the pursuit of broader social equity. Study 1 can be interpreted within the framework of procedural justice, illustrating how user experiences of platform legitimacy are shaped by the perceived fairness of moderation systems and processes. Study 2 draws on retributive justice to explain why certain forms of online abuse are perceived as justified or even deserved, highlighting the need for platform designs that integrate more restorative approaches for addressing and preventing harm. Study 3 extends these insights by revealing how the structural conditions of platform governance—shaped by centralized authority, labor precarity, and other economic incentives—compromise user trust and safety, underscoring the importance of distributive justice in producing more equitable governance outcomes. Collectively, these studies indicate that a justice-oriented approach to social media governance requires not only the reduction of harm, but also fundamental transformation of the underlying structures, processes, and systems that determine how platforms are governed.

6.1 Transforming social media governance

Broadly, this dissertation demonstrates that social media governance is very broken—primarily because it does not account for the structural conditions of harm. Classification systems fail to account for dynamics of power and oppression (Study 1); platform designs do not allow communities to pursue accountability or otherwise reaffirm their values in ways that do not reinforce or perpetuate violence (Study 2); and numerous structural barriers, including capitalist logics of innovation and growth, prevent workers from producing more equitable governance, with even evidence-based improvements repeatedly ignored (Study 3). Because of these structural limitations, truly transforming social media governance requires systemic intervention.

One potential source of inspiration for achieving more equitable governance is *transformative justice*, which aims to address and prevent violence by transforming the specific social conditions that create and perpetuate injustice (Mingus, 2019; Kaba, 2021). While frameworks of procedural, retributive, restorative, and distributive justice offer useful lenses for understanding and evaluating current platform governance practices, transformative justice provides a more radical foundation for reimagining what social media governance could become. Transformative justice emerged as a response to the failures of state systems to provide safety or justice for marginalized communities, instead perpetuating violence through carceral logics of surveillance and punishment (Mingus, 2019; Kaba, 2021). Rooted in abolitionist, Indigenous, and Black feminist traditions, transformative justice seeks not to punish wrongdoing, but to address its underlying causes, with the ultimate goal of creating conditions under which harm is less likely to occur.

By examining the structural conditions that create and perpetuate violence—including racism, transphobia, poverty, and other intersecting systems of oppression (Crenshaw, 1991)—transformative justice invites us to imagine alternative, community-driven futures, reducing reliance on carceral systems by encouraging communities to take collective responsibility for addressing and preventing harm (Mingus, 2019; Kaba, 2021). To cultivate the conditions necessary for healing and safety from harm, transformative justice emphasizes the importance of building relationships, as collective liberation requires a foundation of mutual trust, care, and responsibility. As community organizer Mia Mingus (2019) writes, "violence is collectively enabled, has a collective impact, and requires a collective response."

Applying a transformative justice lens to social media governance requires a fundamental reimagining of how we define and respond to online harm—not just at the level of content and conduct, but structurally. Rather than regarding harm as an isolated violation of platform rules, to be adjudicated by distant moderators or opaque algorithms, a transformative justice approach to platform governance would involve confronting the specific local conditions that allow harm to occur. In practice, this would mean shifting away from punitive, reactive, and top-down enforcement models in favor of community-led governance and structural prevention, including platform designs that mitigate—rather than exacerbate—harm. Achieving this vision of platform governance would require establishing platform policies that actively account for systemic inequity, rather than "neutral" policies that reflect and reproduce existing systems of oppression. It would also require significant investment in preventative infrastructure, such as onboarding experiences that establish shared norms, and community accountability mechanisms that prioritize healing and repair over punishment or exclusion. Ultimately, a transformative justice approach would treat governance not as a matter of control or compliance, but as an ongoing, relational practice of care, repair, and collective responsibility.

Under platform capitalism, such a transformation is unlikely to occur (Srnicek, 2017; Musgrave, 2022). While contemporary platforms rely on punishment, surveillance, and other carceral logics (Zuboff, 2019; York, 2021), transformative justice frameworks reject the notion that safety is produced by policing, instead articulating a vision of abolition grounded in community accountability and care (Mingus, 2019; Kaba, 2021). Advancing more equitable infrastructure requires "coalitional effort" (Irani et al., 2025), and enacting just governance—online and offline—will require sustained commitment to addressing broader social inequities, including dismantling intersecting systems of oppression such as transphobia, white supremacy, colonialism, and other forms of sexism, racism, and imperialism. Although broader social justice movements extend beyond the scope of this dissertation, I conclude with three potential strategies for transforming social media governance under the current realities of platform capitalism: resisting scale, decentralizing authority, and rehumanizing technology.

6.1.1 Resisting scale

Across all three studies in this dissertation, participants described myriad challenges arising from governance systems designed for global scale. Scaled content moderation requires uniformity, but social, political, and linguistic diversity cannot be standardized—resulting in platform governance practices that reproduce existing cultural hegemony, privileging the experiences and needs of dominant groups and reinforcing systemic oppression (Blackwell et al., 2017; Shahid & Vashistha, 2023). Massive volumes of individual enforcement decisions are made by precarious workers who lack sufficient time, tools, and context to appropriately adjudicate nuanced interactions (Roberts, 2016; Roberts, 2019), producing inconsistent, opaque, and often unjust outcomes that disproportionately impact marginalized users (Haimson et al., 2021; Mayworm et al, 2024; Thach et al., 2024). As technology companies struggle to effectively govern their platforms at ever-expanding scales (Gillespie, 2018), users increasingly take justice into their own hands, resorting to public shaming and other punitive forms of retributive harassment that exacerbate, rather than reduce, harm (Blackwell et al., 2018).

Contemporary platform governance is deeply intertwined with the political economy of scalability. Under platform capitalism, a desirable technological solution is one that can be efficiently deployed across billions of users with minimal additional labor, a logic rooted in the continual accumulation of capital (Tsing, 2012; Hanna & Park, 2020). Pfotenhauer et al. (2022) describe this obsession with scalable platform technologies as a “scalability zeitgeist,” under which the promise of expansion obscures the social, political, and ethical consequences of extending scaled solutions across divergent contexts. In the context of social media governance, scale is not merely a technical feature but a political and economic imperative, requiring global standardization and increasingly automated enforcement to achieve efficiency and profitable growth—favoring the interests of investors over users. The purpose of a system is what it does (Beer, 2002), and it is increasingly clear that the purpose of contemporary social media governance is to serve corporate interests by maintaining the racialized, gendered, and colonial logics that sustain capitalist extraction—at the expense of safety, equity, and justice.

Transformative justice is fundamentally incompatible with the logic of scale. As Anna Tsing (2012) cautions, “scalable projects are those that can expand without changing,” while transformative justice frameworks demand exactly the opposite: structural transformation of the social, institutional, and material conditions that create and perpetuate systemic harm. Structural injustice is necessarily ignored by scalable content moderation systems, which rely on binary classifications and similarly reductive models of user behavior that flatten, omit, or otherwise obscure the complexity of harm (Blackwell et al., 2017; Noble, 2018; Benjamin, 2019). Instead, platforms expend considerable resources identifying and responding to isolated incidents of harm

(Gillespie, 2018; Roberts, 2019), divorced from their structural foundations and other essential context required to effectively intervene. In the absence of meaningful repair, unmitigated harms compound, further entrenching the same structural conditions and reinforcing cycles of violence (Schoenebeck & Blackwell, 2021).

Ultimately, transforming social media governance requires challenging these structural disincentives, which may necessarily involve designing platform alternatives that prioritize equity, care, and collective liberation over profitability, efficiency, and scale. Potential alternatives would need to actively oppose dominant market logics, resisting universal, one-size-fits-all solutions in favor of more deliberate, community-driven governance practices. Platform alternatives that offer modular or otherwise adaptable moderation systems could facilitate decentralized governance across a range of pluralistic communities, reimagining what it means to scale.

6.1.2 Decentralizing authority

The limits of scalable governance underscore the need for smaller-scale alternatives that are responsive to the specific needs of diverse user communities. The studies in this dissertation advocate for more participatory approaches to platform design, centering the perspectives of vulnerable users to produce moderation systems that more appropriately account for diverse experiences of harm (Blackwell et al., 2017). In contrast to top-down models of scaled platform governance that prioritize efficiency and abstraction, a transformative justice approach to social media governance would instead decentralize authority, empowering communities of users to determine collective governance practices that align with their specific needs and values. Crucially, decentralization would not mean the absence of structure, but the redistribution of power to users who are historically excluded from platform governance decisions—especially those communities most harmed by platform indifference, inaction, or overreach.

Contemporary platform governance operates through centralized regimes of control, in which a small set of corporate actors define, enforce, and interpret rules for billions of users. Centralized systems are often unable or unwilling to recognize the structural complexity of harm, in part because central authority is created and maintained by the same systems of oppression that produce and perpetuate violence (Mingus, 2019; Kaba, 2021). As this dissertation demonstrates, centralized content moderation systems are similarly structurally incapable of accounting for the relational, contextual, and political complexity of online harm. In a platform governance context, decentralizing authority requires relocating the power to define rules, interpret harm, and enact accountability, shifting authority away from distant, opaque institutions and toward the communities most directly impacted by governance outcomes.

A rich history of HCI and CSCW scholarship has explored online models of distributed governance. On collaborative encyclopedia Wikipedia, users engage in and contribute to both formal and informal decentralized governance processes (Forte et al, 2009; Geiger & Ribes, 2010). Social news website Slashdot distributes moderation responsibilities by allowing users to upvote or downvote individual comments, with only comments above a certain vote threshold displayed (Lampe & Resnick, 2004; Lampe et al., 2014). On Reddit (Jhaver et al., 2019b; Matias, 2019) and Discord (Hwang et al., 2024; Yoon et al., 2025), users join individual subreddits or servers, respectively, each with local rules enforced by community volunteers. Similarly, federated social media platforms such as Mastodon allow users to create and moderate their own servers, each with local rules and the ability to determine which other servers their users can connect or "federate" with (Zignani et al., 2018; Zhang et al., 2024). These and similar platforms demonstrate that decentralized governance is not only possible, but often more participatory, responsive, and context-sensitive than centralized alternatives.

Social media users clearly have desire for more community-driven governance mechanisms, as evidenced by the use of existing decentralized platforms as well as popular self-moderation tools such as blocklists, which allow users to collaboratively define lists of users whose content they wish to block (Geiger, 2016; Jhaver et al., 2018). Yet, as Seering (2020) notes, no major social media platforms incorporate features for facilitating collective moderation decisions, such as moderator election tools or voting mechanisms for potential changes in community rules. Even platforms primarily sustained by volunteer moderation labor exhibit "common tendencies toward oligarchy," with control over community governance held by small groups of individual moderators whose goals may diverge from broader community interests—resulting in governance practices that may not appropriately reflect local values and needs (Matias, 2019). While this dissertation primarily examines the scaled content moderation practices of large, corporate-owned social media platforms, many smaller online communities already self-govern successfully, and have since the early days of the Internet—and indeed, scholars increasingly posit that realizing more equitable online futures may require returning to a broader landscape of smaller, more distributed communities (Fiesler et al., 2018; Matias, 2019; Costanza-Chock, 2020; Seering, 2020; Hasinoff & Schneider, 2022).

Fanfiction community Archive of Our Own (AO3)—which boasts over two million users and more than five million works of fanfiction—was built primarily by volunteers, with a variety of custom-designed writing and tagging features informed by the specific needs and interests of the fanfiction community (Fiesler et al., 2016). Fanfiction authors' long-standing practice of crediting their influences, for example, informed AO3's "citation" feature, which allows authors to directly credit work that inspired their writing (Fiesler & Bruckman, 2019). Embedding AO3's community norms directly into the technical design of the platform is made possible by the designers' proximity to their community, reflecting the localized knowledge required to effectively situate technologies in the specific context of the communities they serve (Haraway, 1991; Suchman, 2002). As corporate social media platforms pursue astronomical scale, they must also pursue increasingly automated governance mechanisms, further distancing content moderation policies and processes from the individual users and communities they purportedly seek to govern.

6.1.3 Rehumanizing technology

Because harm is collectively enabled, it requires a collective response, and transformative justice asserts that only communities have the requisite contextual knowledge to effectively understand and address the causes and consequences of harm (Mingus, 2019; Kaba, 2021). This treatment positions harm not as an abstract violation of arbitrary rules, but as a breakdown in relationships, shaped by intersecting systems like racism, patriarchy, colonialism, and capitalism (Crenshaw, 1991). The community, then, becomes not only the location of harm, but a potential site of healing, repair, and transformation. Exploratory research has demonstrated initial support for these same principles in a social media context, with both users and moderators expressing interest in—and reporting preliminary success with—governance processes designed to remediate harm by promoting accountability, including intervening in potential conflict, facilitating apologies or other forms of redress, and clarifying the specific impacts of harmful behaviors (Jhaver et al., 2019b; Warzel, 2019; Schoenebeck et al., 2021; Xiao et al., 2023; Doan & Seering, 2025). While promising, these approaches remain largely unsupported by dominant corporate platforms, which continue to invest in governance systems that are structurally incompatible with social justice.

The studies in this dissertation reveal that successful social media governance requires steady investment in the relational practices which sustain our individual and collective lives. While often rendered invisible and structurally undervalued, caregiving, cooperation, education, mediation, and other everyday demonstrations of human interdependence are the necessary preconditions of social well-being (Fisher & Tronto, 1990; Fraser, 2016). Contemporary platforms, however, are structured around logics of punishment, profit, and control—actively eroding these same practices. Complex harms are collapsed into binary categories, marginalizing difference (Study 1); accountability is supplanted by vigilantism, chilling public expression (Study 2); and governance labor is outsourced to increasingly precarious workers, fragmenting critical infrastructures of care (Study 3). To rehumanize technology, then, is to reverse these dynamics, in the hopes of prefiguring the necessary conditions for broader social justice.

In practice, this means prioritizing platforms that embed explicit care infrastructure. HeartMob offers one blueprint: designed in direct collaboration with targets of online harassment, HeartMob affords users the agency to request particular forms of social and material support, based on individual needs, preferences, and circumstances. In turn, bystanders are afforded access to clear avenues for demonstrating care and concern, while other product features purposefully reduce the effort typically required to report, document, or condemn identified harms. By facilitating timely access to tangible community support, HeartMob transforms diffuse networks of concern into concrete expressions of care, directly enabling the forms of social reproduction necessary to sustain people, communities, and institutions. These design choices exemplify the same core commitments of transformative justice: centering the needs of those most directly impacted by harm; positioning safety as a collective responsibility; and relocating governance power from distant institutions to communities themselves (Mingus, 2019; Kaba, 2021).

In order for online platforms to ever successfully promote—rather than impair—equity, expression, and human dignity, social media must be rooted in similar commitments. Principles of *design justice,* or the intentional examination of who benefits from (and who is burdened by) specific design choices, offer practical pathways toward the design of novel platforms that recognize difference, resource care, and meaningfully redistribute power (Costanza-Chock, 2020). For example, platform cooperatives, owned and governed by users, could enable communal governance decisions that reflect collective deliberation, rather than fiduciary responsibilities. As demonstrated by the findings of this dissertation, even seemingly small interventions—such as a single visible act of dissent (Study 2)—can dramatically shift normative expectations, evidence of design's ability to more appropriately scaffold the practices required to maintain successful communities. Platform designs that localize authority, facilitate mutual aid, and promote other socially reproductive functions could transform our current understanding of community governance, creating the necessary infrastructure to dismantle structural violence and produce more equitable sociotechnical futures.

6.2 Conclusion

This dissertation advances a justice-oriented framework for understanding and transforming contemporary social media governance. Three empirical research studies demonstrate that current platform governance practices often exacerbate the very harms they seek to mitigate, due in large part to the influence of punitive criminal justice models that are structurally incapable of producing equitable governance outcomes. A justice-oriented approach reframes social media governance from a logic of control to a practice of collective care, understanding harm not as a violation of rules, but as a signal of relational and structural breakdown requiring systemic attention.

Across the three studies in this dissertation, there is clear evidence that harm on social media platforms is not simply the result of individual misbehavior—it is produced and exacerbated by platform designs that incentivize engagement over safety, standardization over nuance, and global expansion over local consideration. Classification systems flatten complex human experiences into binary categories, resulting in enforcement practices that reproduce existing structural oppression (Blackwell et al., 2017). Bystander dynamics reveal the public's inclination toward punishment, rather than repair, perpetuating cycles of violence (Blackwell et al., 2018). And the platform governance ecosystem itself—marked by labor exploitation, institutional opacity, and corporate concentration of power—reflects a broader political economy that treats safety and justice as secondary to profitability (Blackwell, 2025).

Rather than merely refining classification systems or improving moderation workflows, transformative justice invites us to reimagine how online platforms can be restructured to prioritize collective accountability, care, and repair. Crucially, such an approach demands not only better policies and tools, but a fundamental restructuring of the systems and logics that currently govern online spaces. Under this model, governance is not an afterthought, but a core social and political function of the platform itself, requiring meaningful transparency, community participation, and sustained commitment to justice. Achieving this vision requires resisting the imperative to scale and redistributing governance authority to users themselves, particularly those most marginalized by existing platforms.

Applying transformative justice to social media governance also requires platforms to reckon with their role in upholding broader social injustice. Many of the harms discussed across this dissertation are not simply content moderation problems, but manifestations of existing systems of social oppression—racism, misogyny, transphobia, and so on—that platforms reflect, amplify, and entrench. Truly transforming online governance requires attending to global power asymmetries, including histories of oppression, to produce platforms that are appropriately responsive to the diverse needs, values, and identities of the communities they serve. Transformative justice provides a framework both for critique and improvement, offering a vision of governance grounded in mutual accountability and care. The findings of this dissertation demonstrate that such a transformation is not only necessary but possible, reflected in the practices of users, communities, and workers who resist dominant systems by imagining alternative social and technology futures.

References

Ahmed, S. (2017). *Living a Feminist Life.* Duke University Press.

Angwin, J., & Grassegger, H. (2017). *Facebook's Secret Censorship Rules Protect White Men From Hate Speech but not Black Children.* ProPublica.

Ashktorab, Z., & Vitak, J. (2016). Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3895–3905.

Avle, S., Lin, C., Hardy, J., & Lindtner, S. (2020). Scaling Techno-Optimistic Visions. *Engaging Science, Technology, and Society, 6*, 237–254.

Bailey, A. H., & LaFrance, M. (2017). Who Counts as Human? Antecedents to Androcentric Behavior. *Sex Roles, 76*(11), 682–693.

Balkin, J. M. (2021). How to Regulate (and Not Regulate) Social Media. *Journal of Free Speech Law, 1*, 71–96.

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining Interventions to Reduce the Spread of Viral Misinformation. *Nature Human Behaviour, 6*(10), 1372–1380.

Bandura, A. (1971). *Social Learning Theory.* General Learning Press.

Barbrook, R., & Cameron, A. (1996). The Californian Ideology. *Science as Culture, 6*(1), 44–72.

Bardzell, S. (2010). Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1301–1310.

Bardzell, S., & Bardzell, J. (2011). Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 675–684.

Basic Online Safety Expectations (2024, January). *Summary of response from X Corp. (Twitter) to eSafety's transparency notice on online hate.* http://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices

Becerra, L. D., & Thomas, H. E. (2023). Innovation Doesn't Work: The Explanatory Power of a Socio-Technical Approach. *Engaging Science, Technology and Society 9*(2), 66–74.

Becker, H. (1963). *Outsiders.* Free Press.

Beer, S. (2002). What is Cybernetics? *Kybernetes, 31*(2), 209-219.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity.

Blackwell, L. (2025). *Content Moderation Futures.* Unpublished manuscript.

Blackwell, L., Hardy, J., Ammari, T., Veinot, T., Lampe, C., & Schoenebeck, S. (2016). LGBT Parents and Social Media: Advocacy, Privacy, and Disclosure During Shifting Social Movements. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 610–622.

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW).

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When Online Harassment is Perceived as Justified. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

Bogardus, E. S. (1933). A Social Distance Scale. *Sociology & Social Research.*

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences.* The MIT Press.

Boyd, J. (2002). In Community We Trust: Online Security Communication at eBay. *Journal of Computer-Mediated Communication, 7*(3).

Braithwaite, J. (2002). *Restorative Justice & Responsive Regulation.* Oxford University Press.

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Bruckman, A., Danis, C., Lampe, C., Sternberg, J., & Waldron, C. (2006). Managing Deviant Behavior in Online Communities. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 21–24.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work.*

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls Just Want to Have Fun. *Personality and Individual Differences, 67,* 97–102.

Caplan, R. (2018). *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches.* Data and Society Institute.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.

Carlsmith, K. M., & Darley, J. M. (2008). Psychological Aspects of Retributive Justice. *Advances in Experimental Social Psychology, 40*, 193–236.

Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need.* The MIT Press.

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1201–1213.

Chander, A., & Krishnamurthy, V. (2018). The Myth of Platform Neutrality. *Georgetown Law Technology Review, 2*, 400–416.

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017). 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*, 3175–3187.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., . . . Gilbert, E. (2018). The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW).

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230.

Cho, D., & Acquisti, A. (2013). The More Social Cues, the Less Trolling? An Empirical Study of Online Commenting Behavior. In *Proceedings of the 12th Workshop on the Economics of Information Security*.

Chung, A., & Rimal, R. N. (2016). Social Norms: A Review. *Review of Communication Research, 4*, 1–28.

Cialdini, R. B. (2001). *Influence: Science and Practice.* Allyn and Bacon.

Cialdini, R. B. (2007). Descriptive Social Norms as Underappreciated Sources of Social Control. *Psychometrika 72*(2).

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Citron, D. K., & Franks, M. A. (2014). Criminalizing Revenge Porn. *Wake Forest Law Review, 49*.

Citron, D. K., & Franks, M. A. (2020). The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform. *University of Chicago Legal Forum,* 45–75.

Clark, K. (2015). Online Violence Against Trans Women Perpetuates Dangerous Cycle. *The Huffington Post*.

Clarke, A. E. (2005). *Situational Analysis: Grounded Theory After the Postmodern Turn.* Sage Publishing.

Collins, K. (2017). Tech is Overwhelmingly White and Male, and White Men are Just Fine With That. *Quartz.*

Collins, P. H. (1990). *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*. Hyman.

Crawford, K., & Gillespie, T. (2016). What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society, 18*(3), 410–428.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review, 43*(6), 1241–1299.

Cryst, E., Grossman, S., Hancock, J., Stamos, A., & Thiel, D. (2021). Introducing the Journal of Online Trust and Safety. *Journal of Online Trust and Safety, 1*(1).

Darley, J. M., & B. Latané. (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology 8*(4): 377–383.

DeVito, M. A., Walker, A. M., & Fernandez, J. R. (2021). Values (Mis)alignment: Exploring Tensions Between Platform and LGBTQ+ Community Design Values. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW).

Diakopoulos, N., & Naaman, M. (2011). Towards Quality Discourse in Online News Comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 133–142.

Dibbell, J. (1993). A Rape in Cyberspace: How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database Into a Society. *The Village Voice*.

Dibbell, J. (1998). *My Tiny Life: Crime and Passion in a Virtual World.* Henry Holt & Company.

Dillon, K. P., & Bushman, B. J. (2015). Unresponsive or Unnoticed?: Cyberbystander Intervention in an Experimental Cyberbullying Context. *Computers in Human Behavior, 45*, 144–150.

Dimond, J. P., Dye, M., LaRose, D., & Bruckman, A. S. (2013). Hollaback! The Role of Storytelling Online in a Social Movement Organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 477–490.

Doan, B. N., & Seering, J. (2025). The Design Space for Online Restorative Justice Tools: A Case Study with ApoloBot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Dombrowski, L., Harmon, E., & Fox, S. (2016). Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 656–671.

Dombrowski, L., Alvarado Garcia, A., & Despard, J. (2017). Low-Wage Precarious Workers' Sociotechnical Practices Working Towards Addressing Wage Theft. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4585–4598.

Domínguez Hernández, A., Ramokapane, K. M., Das Chowdhury, P., Michalec, O., Johnstone, E., Godwin, E., . . . Rashid, A. (2023). Co-Creating a Transdisciplinary Map of Technology-Mediated Harms, Risks and Vulnerabilities: Challenges, Ambivalences and Opportunities. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW).

Donath, J. S. (1999). Identity and Deception in the Virtual Community. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace*, 37–68. Routledge.

Douek, E. (2020). Australia's “Abhorrent Violent Material” Law: Shouting “Nerd Harder” and Drowning Out Speech. *Australian Law Journal, 94*(1), 41–60.

Dourish, P., & Bell, G. (2011). *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing.* The MIT Press.

Dourish, P., & Mainwaring, S. D. (2012). Ubicomp's Colonial Impulse. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 133–142.

Duff, A. S. (2016). Rating the Revolution: Silicon Valley in Normative Perspective. *Information, Communication & Society, 19*(11), 1605–1621.

Duggan, M., Rainie, L., Smith, A., Funk, C., Lenhart, A., & Madden, M. (2014). *Online Harassment.* Pew Research Center.

Duggan, M., & Smith, A. (2017). *Online Harassment*. Pew Research Center.

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine. *Convergence, 26*(2), 237–252.

Durkheim, E. (1884). The Division of Labor in Society. *Journal des Économistes*.

Ehrenreich, B., & Ehrenreich, J. (1977). The Professional-Managerial Class. *Radical America,* 11(2), 7–32.

Erikson, K. (1966). *Wayward Puritans.* Wiley.

Fiesler, C., Morrison, S., & Bruckman, A. S. (2016). An Archive of Their Own: A Case Study of Feminist HCI and Values in Design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2574–2585.

Fiesler, C., Jiang, J., McCann, J., Frye, K., & Brubaker, J. (2018). Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

Fiesler, C., & Bruckman, A. S. (2019). Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement. *Proceedings of the ACM on Human-Computer Interaction, 3*(GROUP).

Fisher, B. & Tronto, J. C. (1990). Toward a Feminist Theory of Caring. In E. K. Abel & M. K. Nelson (Eds.), *Circles of Care: Work and Identity in Women's Lives*, 36–54. State University of New York Press.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., . . . Kainbacher, M. (2011). The Bystander-Effect: a Meta-Analytic Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies. *Psychological Bulletin 137*(4).

Forte, A., Larco, V., & Bruckman, A. (2009). Decentralization in Wikipedia Governance. *Journal of Management Information Systems, 26*(1), 49–72.

Foucault, M., & Nazzaro, A. M. (1972). History, Discourse and Discontinuity. *Salmagundi, 20*, 225–248.

Fox, S. E., Khovanskaya, V., Crivellaro, C., Salehi, N., Dombrowski, L., Kulkarni, C., . . . Forlizzi, J. (2020). Worker-Centered Design: Expanding HCI Methods for Supporting Labor. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

Fraser, N. (2016). Contradictions of Capital and Care. *New Left Review, 100*(99).

Garfinkel, H. (1967). *Studies in Ethnomethodology.* Prentice-Hall.

Geiger, R. S. (2016). Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space. *Information, Communication & Society, 19*(6), 787–803.

Geiger, R. S., & Ribes, D. (2010). The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 117–126.

Gillespie, T. (2010). The Politics of 'Platforms'. *New Media & Society, 12*(3), 347–364.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press.

Giner-Sorolla, R., Bosson, J., Caswell, A., & Hettinger, V. (2012). Emotions in Sexual Morality: Testing the Separate Elicitors of Anger and Disgust. *Cognition & Emotion 26*(7): 1208–1222.

Goldman, E. (2020). *A Pre-History of the Trust & Safety Professional Association (TSPA).* Technology & Marketing Law Blog. http://blog.ericgoldman.org/archives/2020/06/a-pre-history-of-the-trust-safety-professional-association-tspa.htm

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research 35*(3), 472–482.

Goode, E., & Ben-Yehuda, N. (2009). *Moral Panics: The Social Construction of Deviance.* Wiley-Blackwell.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society, 7*(1).

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley From Building a New Global Underclass.* Harper Business.

Griffiths, M. (2002). Occupational Health Issues Concerning Internet Use in the Workplace. *Work & Stress, 16*(4), 283–286.

Grimmelmann, J. (2013). Speech Engines. *Minnesota Law Review, 98*, 868–952.

Grimmelmann, J. (2014). *Internet Law: Cases and Problems 4.0.* Semaphore Press.

Haimson, O. L., & Hoffmann, A. L. (2016). Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities. *First Monday, 21*(6).

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW).

Han, C., Seering, J., Kumar, D., Hancock, J. T., & Durumeric, Z. (2023). Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW).

Hanna, A., & Park, T. M. (2020). Against Scale: Provocations and Resistances to Scale Thinking. arXiv preprint arXiv:2010.08850.

Haraway, D. (1991). *Simians, Cyborgs, and Women.* Routledge.

Hardaker, C. (2010). Trolling in Asynchronous Computer-Mediated Communication: From User Discussions to Academic Definitions. *Journal of Politeness Research, 6*(2).

Harding, S. (1992). Rethinking Standpoint Epistemology: What is "Strong Objectivity?" *The Centennial review, 36*(3), 437–470.

Hardy, J. (2019). How the Design of Social Technology Fails Rural America. In *Companion Publication of the 2019 Designing Interactive Systems Conference*, 189–193.

Hardy, J., Geier, C., Vargas, S., Doll, R., & Howard, A. L. (2022). LGBTQ Futures and Participatory Design: Investigating Visibility, Community, and the Future of Future Workshops. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW).

Hasinoff, A., Gibson, A. & Salehi, N. (2020). *Restorative Justice for Addressing Online Harm.* Brookings Institution.

Hasinoff, A. A., & Schneider, N. (2022). From Scalability to Subsidiarity in Addressing Online Harm. *Social Media + Society, 8*(3).

Hechter, M., & Opp, K. D. (Eds.). (2001). *Social Norms.* Russell Sage Foundation.

Helles, R., & Lomborg, S. (2024). Techlash or Tech Change? How the Image of Mark Zuckerberg Changed with Cambridge Analytica. In K. Albris, K. Fast, F. Karlsen, A. Kaun, S. Lomborg, & T. Syvertsen (Eds.), *The Digital Backlash and the Paradoxes of Disconnection*, 25–43. Nordicom.

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., & Burghardt, K. (2023). Auditing Elon Musk's Impact on Hate Speech and Bots. In *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 1133–1137.

Hooks, B. (2000). *Feminism is for Everybody: Passionate Politics.* Pluto Press.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv preprint arXiv:1702.08138.

Husovec, M. (2023). Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules. *Berkeley Technology Law Journal, 38*.

Hwang, S., Kiene, C., Ong, S., & Shaw, A. (2024). Adopting Third-Party Bots for Managing Online Communities. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW).

Im, J., Schoenebeck, S., Iriarte, M., Grill, G., Wilkinson, D., Batool, A., . . . Naseem, M. (2022). Women's Perspectives on Harm and Justice After Online Harassment. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW).

Irani, L. (2015a). Difference and Dependence Among Digital Workers: The Case of Amazon Mechanical Turk. *South Atlantic Quarterly, 114*(1), 225–234.

Irani, L. (2015b). Hackathons and the Making of Entrepreneurial Citizenship. *Science, Technology, & Human Values, 40*(5), 799–824.

Irani, L. (2019). *Chasing Innovation: Making Entrepreneurial Citizens in Modern India.* Princeton University Press.

Irani, L. (2023). Encountering Innovation, Countering Innovation. *Engaging Science, Technology, and Society, 9*(2), 118–130.

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620.

Jackson, S. J. (2014). Rethinking Repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society,* 221–239. The MIT Press.

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI), 25*(2).

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019a). “Did You Suspect the Post Would be Removed?” Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on human-computer interaction, 3*(CSCW).

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019b). Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI), 26*(5).

Jhaver, S., Bruckman, A., & Gilbert, E. (2019c). Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW).

Jiang, J. A., Middler, S., Brubaker, J. R., & Fiesler, C. (2020). Characterizing Community Guidelines on Social Media Platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, 287–291.

Jungk, R., & Müllert, N. (1987). *Future Workshops: How to Create Desirable Futures.* Institute for Social Inventions.

Kaba, M. (2021). *We Do This ‘Til We Free Us: Abolitionist Organizing and Transforming Justice.* Haymarket Books.

Kant, I. (1911). *The Critique of Judgement* (J. C. Meredith, Trans.). Clarendon Press. (Original work published 1781)

Katsaros, M., Tyler, T., Kim, J., & Meares, T. (2022). Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. *Journal of Online Trust and Safety, 1*(3).

Katsaros, M., Kim, J., & Tyler, T. (2024). Online Content Moderation: Does Justice Need a Human Face? *International Journal of Human–Computer Interaction, 40*(1), 66–77.

Kawakita, H., Nishimura, M., Satoh, Y., & Shibata, N. (1967). Neurological Aspects of Behçet's Disease: A Case Report and Clinico-Pathological Review of the Literature in Japan. *Journal of the Neurological Sciences, 5*(3), 417–439.

Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying Bystander Intervention: The Number of Offenders and Retweeting Predict Likelihood of Helping a Cyberbullying Victim. *Journal of Computer-Mediated Communication, 23*(3), 146–162.

Keller, D. (2022). *The EU’s new Digital Services Act and the Rest of the World.* Verfassungsblog.

Keller, D. (2023). Platform Transparency and the First Amendment. *Journal of Free Speech Law, 4*.

Keller, D., & Leerssen, P. (2020). Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In N. Persily & J. Tucke (Eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform,* 220–251. Cambridge University Press.

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating Behavior in Online Communities. In R. E. Kraut & P. Resnick (Eds.), *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Klonick, K. (2017). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review, 131*, 1598–1670.

Klonick, K. (2019). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *Yale Law Journal, 129*, 2418–2499.

Knittel, M., & Menking, A. (2024). Bridging Theory & Practice: Examining the State of Scholarship Using the History of Trust and Safety Archive. *Journal of Online Trust and Safety, 2*(2).

Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J., & Kiesler, S. (1996). HomeNet: A Field Trial of Residential Internet Services. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 284–291.

Kraut, R. E., & Resnick, P. (Eds.) (2012). Building Successful Online Communities: Evidence-Based Social Design. MIT Press.

Lampe, C., & Resnick, P. (2004). Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550.

Lampe, C., & Johnston, E. (2005). Follow the (Slash)dot: Effects of Feedback on New Members in an Online Community. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*, 11–20.

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010). Motivations to Participate in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Lampe, C. (2014). Gamification and Social Media. In S. P. Walz & S. Deterding (Eds.), *The Gameful World: Approaches, Issues, Applications*, 461–480. The MIT Press.

Lampe, C., Zube, P., Lee, J., Park, C. H., & Johnston, E. (2014). Crowdsourcing Civility: A Natural Experiment Examining the Effects of Distributed Moderation in Online Forums. *Government Information Quarterly, 31*(2), 317–326.

Lawrence, R. (1991). Reexamining Community Corrections Models. *Crime & Delinquency, 37*(4), 449–464.

Lea, M., & Spears, R. (1991). Computer-Mediated Communication, De-Individuation and Group Decision-Making. *International Journal of Man-Machine Studies, 34*(2), 283–301.

Lee, R. (2020, May). *Tech Layoff Tracker.* Layoffs.fyi. Retrieved July 27, 2025, from http://layoffs.fyi

Lenhart, A., Ybarra, M., Zickuhr, K., & Prive-Feeney, M. (2016). *Online Harassment, Digital Abuse, and Cyberstalking in America.* Data & Society Research Institute.

Lessig, L. (2006). *Code: Version 2.0.* Basic Books.

Litt, E., & Hargittai, E. (2016). The Imagined Audience on Social Network Sites. *Social Media + Society, 2*(1).

Lombroso, C. (1911). *Criminal Man* (G. Lombroso-Ferrero, Trans.). Putnam. (Original work published 1876)

Lyu, Y., Cai, J., Callis, A., Cotter, K., & Carroll, J. M. (2024). "I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Marwick, A. (2012). The Public Domain: Surveillance in Everyday Life. *Surveillance & Society, 9*(4), 378–393.

Marwick, A. E., & Miller, R. (2014). *Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape.* Fordham Center on Law and Information Policy Report.

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online.* Data & Society Research Institute.

Marwick, A., Kuo, R., Cameron, S. J. & Weigel, M. (2021). Critical Disinformation Studies: A Syllabus. *Center for Information, Technology, & Public Life*.

Marx, K. (1959). *Estranged Labor* (M. Mulligan, Trans.). Progress Publishers. (Original work published 1844)

Marx, K. & Engels, F. (1967). *The Communist Manifesto* (S. Moore, Trans.). Penguin. (Original work published 1848)

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures. *New Media & Society, 19*(3), 329–346.

Matias, J. N. (2017). *Posting Rules in Online Discussions Prevents Problems & Increases Participation.* CivilServant.

Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online. *Social Media + Society, 5*(2).

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, Reviewing, and Responding to Harassment on Twitter.* Women, Action, and the Media.

Maxim, K., Parecki, J., & Cornett, C. (2022). How to Build a Trust and Safety Team In a Year: A Practical Guide From Lessons Learned (So Far) at Zoom. *Journal of Online Trust and Safety, 1*(4).

Mayworm, S., DeVito, M. A., Delmonaco, D., Thach, H., & Haimson, O. L. (2024). Content Moderation Folk Theories and Perceptions of Platform Spirit Among Marginalized Social Media Users. *ACM Transactions on Social Computing, 7*.

Merlan, A. (2015). The Cops Don't Care About Violent Online Threats. What Do We Do Now? *Jezebel*.

Meta. (2025a). *More Speech and Fewer Mistakes.* Newsroom. http://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Meta. (2025b). *Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook.* Transparency Center. http://transparency.meta.com/reports/regulatory-transparency-reports/

Mezrich, B. (2023). *Breaking Twitter: Elon Musk and the Most Controversial Corporate Takeover in History.* Pan Macmillan.

Milgram, S., Bickman, L., and Berkowitz, L. (1969). Note on the Drawing Power of Crowds of Different Size. *Journal of Personality and Social Psychology, 13*(2), 79–82.

Mingus, M. (2022). Transformative Justice: A Brief Description. *Fellowship, 84*(2), 17–19.

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior, 39*(3), 629–649.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

Opp, K. D. (2001). How Do Norms Emerge? An Outline of a Theory. *Mind & Society, 2*(1), 101–128.

Palen, L., & Dourish, P. (2003). Unpacking "Privacy" for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 129–136.

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work.*

Park, S., & Sang, Y. (2023). The Changing Role of Nation States in Online Content Governance: A Case of Google's Handling of Government Removal Requests. *Policy & Internet, 15*(3), 351–369.

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, 369–374.

Pater, J., & Mynatt, E. (2017). Defining Digital Self-Harm. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1501–1513).

Pearce, K. E., Vitak, J., & Barta, K. (2018). Socially Mediated Visibility: Friendship and Dissent in Authoritarian Azerbaijan. *International Journal of Communication, 12*, 1310–1331.

Perez, S. (2017). *Twitter Adds More Anti-Abuse Measures Focused on Banning Accounts, Silencing Bullying.* TechCrunch.

Pfotenhauer, S., Laurent, B., Papageorgiou, K., & Stilgoe, A. J. (2022). The Politics of Scaling. *Social Studies of Science, 52*(1), 3–34.

Pittaro, M. L. (2007). Cyber Stalking: An Analysis of Online Harassment and Intimidation. *International Journal of Cyber Criminology, 1*(2), 180–197.

Poor, N. (2005). Mechanisms of an Online Public Sphere: The Website Slashdot. *Journal of Computer-Mediated Communication, 10*(2).

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior. *Personality and Social Psychology Bulletin, 27*(10), 1243–1254.

Prinz, J. (2008). Is Morality Innate. *Moral Psychology, 1*, 367–406.

Puig de La Bellacasa, M. (2011). Matters of Care in Technoscience: Assembling Neglected Things. *Social Studies of Science, 41*(1), 85–106.

Qiwei, L., McDonald, A., Haimson, O. L., Schoenebeck, S., & Gilbert, E. (2024). The Sociotechnical Stack: Opportunities for Social Computing Research in Non-Consensual Intimate Media. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW).

Rainie, L., Anderson, J., & Albright, J. (2017). *The Future of Free Speech, Trolls, Anonymity and Fake News Online.* Pew Research Center.

Rheingold, H. (1991). *Virtual Reality: Exploring the Brave New Technologies.* Simon & Schuster.

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier.* Harper Collins.

Riek, L. D., & Irani, L. (2025). The Future Is Rosie?: Disempowering Arguments About Automation and What to Do About It. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Rifat, M. R., Asha, A. Z., Jadon, S., Yan, X., Guha, S., & Ahmed, S. I. (2024). Combating Islamophobia: Compromise, Community, and Harmony in Mitigating Harmful Online Content. *ACM Transactions on Social Computing, 7*.

Roberts, S. T. (2016). *Commercial Content Moderation: Digital Laborers' Dirty Work.* Media Studies Publications.

Roberts, S. T. (2019). *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation.* Yale University Press.

Ronson, J. (2015). How One Stupid Tweet Blew Up Justine Sacco’s Life. *The New York Times*.

Rosner, D. K., Shorey, S., Craft, B. R., & Remick, H. (2018). Making Core Memory: Design Inquiry Into Gendered Legacies of Engineering and Craftwork. In *Proceedings of the 2018 CHI conference on human factors in computing systems*.

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the Platform: Content Moderators and the Logic of Care. *New Media & Society, 22*(6), 1026–1042.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science, 18*(5), 429–434.

Shachaf, P., & Hara, N. (2010). Beyond Vandalism: Wikipedia Trolls. *Journal of Information Science, 36*(3), 357–370.

Schiffer, Z. (2024). *Extremely Hardcore: Inside Elon Musk's Twitter.* Penguin.

Schoenebeck, S., Ellison, N. B., Blackwell, L., Bayer, J. B., & Falk, E. B. (2016). Playful Backstalking and Serious Impression Management: How Young Adults Reflect on Their Past Identities on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1475–1487.

Schoenebeck, S., & Blackwell, L. (2021). Reimagining Social Media Governance: Harm, Accountability, and Repair. *Yale Journal of Law & Technology, 23*, 113–152.

Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2021). Drawing From Justice Theories to Support Targets of Online Harassment. *New Media & Society, 23*(5), 1278–1300.

Scupin, R. (1997). The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization, 56*(2), 233–237.

Seering, J. (2020). Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW).

Seering, J., Kaufman, G., & Chancellor, S. (2022). Metaphors in Moderation. *New Media & Society, 24*(3), 621-640.

Severance, C. (2013). Mitchell Baker: The Mozilla Foundation. *Computer, 46*(2), 7–9.

Shahid, F., & Vashistha, A. (2023). Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Shear, M. D. (2016). Trump as Cyberbully in Chief? Twitter Attack on Union Boss Draws Fire. *The New York Times*.

Sherif, M. (1936). *The Psychology of Social Norms.* Harper.

Simonson, I., & Staw, B. M. (1992). Deescalation Strategies: A Comparison of Techniques for Reducing Commitment to Losing Courses of Action. *Journal of Applied Psychology, 77*(4), 419–426.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its Nature and Impact in Secondary School Pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376–385.

Sproull, L., & Kiesler, S. (1991). *Connections: New Ways of Working in the Networked Organization.* The MIT Press.

Srnicek, N. (2017). *Platform Capitalism.* Wiley.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science, 19*(3), 387–420.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research, 7*(1), 111–134.

Star, S. L., & Strauss, A. (1999). Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work (CSCW), 8*, 9–30.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW).

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending With Group Image: The Psychology of Stereotype and Social Identity Threat. *Advances in Experimental and Social Psychology, 14*, 379–407.

Sternberg, J. (2012). *Misbehavior in Cyber Places: The Regulation of Online Conduct in Virtual Communities on the Internet.* Rowman & Littlefield.

Su, N. M., Lazar, A., & Irani, L. (2021). Critical Affects: Tech Work Emotions Amidst the Techlash. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW).

Suchman, L. (1995). Making Work Visible. *Communications of the ACM, 38*(9), 56–64.

Suchman, L. (2002). Located Accountabilities in Technology Production. *Scandinavian Journal of Information Systems, 14*(2).

Suler, J. (2004). The Online Disinhibition Effect. *Cyberpsychology & Behavior, 7*(3), 321–326.

Suzor, N. P. (2019). *Lawless: The Secret Rules That Govern our Digital Lives.* Cambridge University Press.

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication, 13*.

Sydell, L. (2017). Kyle Quinn Hid at a Friend's House After Being Misidentified on Twitter as a Racist. *National Public Radio*.

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2024). (In)visible Moderation: A Digital Ethnography of Marginalized Users and Content Moderation on Twitch and Reddit. *New Media & Society, 26*(7), 4034–4055.

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation, 27*(2), 237–246.

Trust & Safety Professional Association (2025). *About Us.* http://www.tspa.org/about-tspa

Tsukayama, H. (2017). Twitter Lost 2 Million Users in the U.S. Last Quarter. *The Washington Post*.

Tyler, T. R. (2006). *Why People Obey the Law.* Princeton University Press.

Tyler, T. R. (2007). Procedural Justice and the Courts. *Court Review: The Journal of the American Judges Association, 217*.

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2021). Social Media Governance: Can Social Media Companies Motivate Voluntary Rule Following Behavior Among Their Users? *Journal of experimental criminology, 17*(1), 109–127.

Tyler, T., Meares, T., & Katsaros, M. (2025). New Worlds Arise: Online Trust and Safety. *Annual Review of Criminology, 8*.

U.S. Department of Labor. (2015). *Current Population Survey: Detailed Occupation by Sex and Race.* Bureau of Labor Statistics.

U.S. Equal Employment Opportunity Commission. (2024, September 10). *High Tech, Low Inclusion: Diversity in the High Tech Workforce and Sector, 2014-2022.*

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW).

Vidal, R. V. V. (2006). The Future Workshop: Democratic Problem Solving. *Economic Analysis Working Papers, 5*(4).

Vitak, J., Blasiola, S., Patil, S., & Litt, E. (2015). Balancing Audience and Privacy Tensions on Social Network Sites: Strategies of Highly Engaged Users. *International Journal of Communication, 9*(20).

Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017). Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1231–1245.

Vogels, E. A. (2021). *The State of Online Harassment.* Pew Research Center.

Walen, A. (2015). Proof Beyond a Reasonable Doubt: A Balanced Retributive Account. *Louisiana Law Review, 76*, 355–446.

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research, 23*(1), 3–43.

Walther, J. B. (2002). Cues Filtered Out, Cues Filtered In: Computer Mediated Communication and Relationships. In M. Knapp & J. A. Daly (Eds.), *Handbook of Interpersonal Communication*. Sage Publications.

Warren, C. (2017). Twitter's New Abuse Filter Works Great, If Your Name Is Mike Pence. *Gizmodo*.

Warzel, C. (2019). Could Restorative Justice Fix the Internet? *New York Times*.

Weeks, J. (1999). Discourse, Desire and Sexual Deviance: Some Problems in a History of Homosexuality. In R. G. Parker & P. Aggleton (Eds.), *Culture, Society and Sexuality: A Reader*, 125–149.

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2007). Retributive and Restorative Justice. *Law and Human Behavior, 32*(5), 375–389.

West, S. M. (2018). Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms. *New Media & Society, 20*(11), 4366–4383.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems.* AI Now.

Williams, A. M., & Irani, L. (2010). There's Methodology in the Madness: Toward Critical HCI Ethnography. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2725–2734.

Williams, R. (2014). Facebook's 71 Gender Options Come to UK Users. *The Telegraph*.

Wolf, C. T., Asad, M., & Dombrowski, L. S. (2022). Designing Within Capitalism. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 439–453.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*.

Xiao, S., Jhaver, S., & Salehi, N. (2023). Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Transactions on Computer-Human Interaction, 30*(6).

Yamamoto, S. (2014). *The Reasons We Punish: Creating and Validating a Measure of Utilitarian and Retributive Punishment Orientation* [Master's Thesis, Carleton University]. Carleton University Institutional Repository.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of Harassment on Web 2.0. *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0) Workshop at WWW2009*.

Yoon, J., Zhang, A. X., & Seering, J. (2025). “It's Great Because It's Ran By Us”: Empowering Teen Volunteer Discord Moderators to Design Healthy and Engaging Youth-Led Online Communities. *Proceedings of the ACM on Human-Computer Interaction, 9*(CSCW).

Zhang, Z., Zhao, J., Wang, G., Johnston, S. K., Chalhoub, G., Ross, T., . . . Shadbolt, N. (2024). Trouble in Paradise? Understanding Mastodon Admin's Motivations, Experiences, and Challenges Running Decentralised Social Media. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW).

Zignani, M., Gaito, S., & Rossi, G. P. (2018). Follow the “Mastodon”: Structure and Evolution of a Decentralized Online Social Network. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 541–550.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs.

Zuckerman, E., & Rajendra-Nicolucci, C. (2023). From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms’ Crisis of Legitimacy. *Social Media + Society 9*(3).