Title

Empowering Online Harm Survivors to Addressing Harm with a Restorative Justice Approach

Author

Xiao, Sijia

Publication Date

2024-10-01

Empowering Online Harm Survivors to Addressing Harm with a Restorative Justice Approach

By

Sijia Xiao

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Niloufar Salehi, Co-chair
Professor Coye Cheshire, Co-chair
Professor Morgan Ames
Professor Amy Bruckman
Professor Kimiko Royokai

Fall 2024

Empowering Online Harm Survivors to Addressing Harm with a Restorative Justice Approach

Abstract

Empowering Online Harm Survivors to Addressing Harm with a Restorative Justice Approach

By

Sijia Xiao

Doctor of Philosophy in Information Science

University of California, Berkeley

Professor Niloufar Salehi, Co-chair

Professor Coye Cheshire, Co-chair

Online interpersonal harm, such as harassment and discrimination, is prevalent on social media platforms. Most platforms adopt content moderation as the primary solution, relying on measures like bans and content removal. These measures follow principles of punitive justice, which holds that perpetrators of harm should receive punishment in proportion to the offense. However, these strategies often fall short of addressing the needs of affected individuals — the survivors — who are typically excluded from decision-making and left with various unmet needs.

My dissertation adopts a restorative justice lens and investigates ways to empower online harm survivors in addressing their unique needs. This approach emphasizes survivors' needs and agency, reconceptualizes views on perpetrator accountability, and mobilizes community resources for a collective response to the issue. Through interviews and co-design sessions with survivors, I identified key survivor needs such as sensemaking, emotional support, safety, retribution, and transformation. Building on these insights, I focused on survivors' needs for sensemaking and developed a social computing system to facilitate a structured sensemaking process, connecting survivors with available resources and stakeholders in addressing the harm. Furthermore, I applied a restorative justice lens to understand how survivors, perpetrators, and moderators currently navigate harm within existing moderation practices, examining both the opportunities and challenges of integrating restorative justice practices into the content moderation landscape.

My research highlights the urgent need for social media platforms to incorporate justice and ethical values into their operational frameworks. It advocates for a survivor-centered approach to addressing online interpersonal harm, viewing it as a multi-stakeholder, cross-platform process. This approach calls for a shift towards viewing harm as a communal

challenge and emphasizes cultivating a resilient community culture in the long term, rather than perceiving harm as isolated incidents within a perpetrator-centric framework.

Contents

Contents **i**

1 Introduction **1**

2 Related Work **4**
2.1 Online Harm and Content Moderation 4
2.2 Restorative Justice 6

3 Positionality Statement **10**

4 RQ1. What do survivors need in addressing online harm? **11**
4.1 Introduction 11
4.2 Method 13
4.3 Findings: Stakeholders and Actions 19
4.4 Findings: Timelines of Needs 26
4.5 Discussion 29
4.6 Conclusion 33

5 RQ2: How do we meet online harm survivors' needs of sensemaking? **34**
5.1 Introduction 34
5.2 System Design 35
5.3 Evaluation 43
5.4 Result 47
5.5 Discussion 56
5.6 Conclusion 61

6 RQ3: What are the opportunities and challenges of implementing restorative justice in the current moderation landscape? **62**
6.1 Introduction 62
6.2 Background 64
6.3 Methods 66
6.4 Findings: Current Moderation Models Through a Restorative Justice Lens . 70
6.5 Findings: Online Restorative Justice Possibilities and Challenges 79

6.6 Discussion 88
6.7 Conclusion 95

7 Discussion and Conclusion **96**
7.1 A Culture of Reparation: Shifting Focus from Rules to Harm 96
7.2 Adapting Offline Justice Models for Online Contexts 98
7.3 Concluding Thoughts: Toward Diverse Justice Approaches in Digital Space . 100

Bibliography **102**

Acknowledgments

I am grateful to my PhD advisors, Niloufar Salehi and Coye Cheshire, for their enduring support, wisdom, and care throughout my PhD journey. You have been integral to my growth and joy in research. I sincerely appreciate my committee members, Morgan Ames, Amy Bruckman, and Kimiko Ryokai, for their guidance, support, and invaluable insights. A special thanks to Amy Bruckman for her consistent advice and support throughout both my master's and PhD. I am also grateful to Steven Dow for introducing me to social computing research and offering me my first research experience in the United States. My heartfelt thanks extend to Xiaoru Yuan, Yuanchun Shi, and Chun Yu for guiding my early research during my undergraduate years.

I have made wonderful friends along the way. I am thankful to my collaborators, Shagun Jhaver, Danaë Metaxa, Joon Sung Park, Haodi Zou, Jingshu Rui, Amy Matthews, Weichen Liu, and Jacob Browne, for being a joy to work with and for sharing their expertise and brilliant ideas. I cherish my labmates — Tonya Nguyen, Liza Gak, Samantha Robertson, Sabriya Alam, Angela Jin, and Seyi Olojo — for their friendship, joy, and solidarity. Furthermore, I am grateful to Emma Lurie, Ji Su Yoo, Suraj Nair, Richmond Y. Wong, Anne Jonas, Elizabeth Resor, Noura Howell, Daniel Griffin, Guanghua Chi, Max T. Curran, Zoe Kahn, Renkai Ma, Wesley Deng, Tzu-Sheng Kuo, Michelle Lam, Joseph Seering, and Deepak Kumar for sharing this journey and inspiring me along the way.

I am grateful to UC Berkeley, particularly the I School, for fostering a strong sense of community. I appreciate the I School faculty for their wisdom and for serving as inspiring role models, especially those I have engaged with or taken classes from, such as Paul Duguid, Deirdre Mulligan, and John Chuang. A special thanks to Inessa Gelfenboym Lee, who always responds promptly and genuinely cares about students' well-being. I also deeply appreciate Julie Shackford-Bradley from the Restorative Justice Center for her generous support and inspiration for my dissertation research.

Finally, I am thankful to my family for their unconditional love and support, which have allowed me to pursue my interests and dreams. I am grateful to my partner Ryan Pham, his family, and the family dog Mako, for their love, care, and for sharing life's moments with me in San Jose. I am thankful for dance, which has been a source of delight, enriching my PhD experience and becoming a lifelong passion. Lastly, I am thankful to myself for continually reflecting, improving, and growing as both a researcher and a person.

Chapter 1

Introduction

The expansion of social media platforms has underscored the escalating challenges of online interpersonal harm such as including cyberbullying and sexual harassment [48, 170]. Nearly 41% American adults have experienced harassment, and severe cases are still on the rise.

Online platforms primarily address these types of harm through content moderation, which focuses on regulating perpetrators through punitive actions such as removal of their content or banning their accounts. However, the affected party in the harm, the survivors, are often out of the picture. In a perpetrator-centered framework, survivors are not involved in the process of addressing harm and suffer from a lack of agency, power, and control in meeting their needs. Research has found that survivors have a range of needs that are not addressed by content moderation, including seeking advice, obtaining emotional support, and receiving acknowledgment of the harm and an apology from the person who caused it [145, 121].

Given the growing scale and ramifications of online harm [48, 170], empowering survivors is a matter of societal and ethical urgency. Empowerment can be seen as the process by which individuals and collectives gain control over issues that affect them [129]. In the context of online harm, empowerment for survivors can be understood as a process that enables them to gain agency, control, and power over addressing the harm they have experienced.

My dissertation research draws from restorative justice – a survivor-centered justice approach – to empower online harm survivors to address their needs. An approach that aligns with punitive justice, such as content moderation, responds to harm by centering the offending party and regulating their offending behavior through punishment. Restorative justice, on the other hand, centers survivors' agency and needs in addressing the harm. Restorative justice achieves survivor empowerment through providing guidance and support to help survivors identify their needs and utilize resources from multiple stakeholders related to the harm to address those needs [184, 3].

Restorative justice philosophy and practices align with the goal of giving online harm survivors agency, control, and power to empower them. It has been successfully applied in a myriad of offline settings that traditionally adopt a holistic punitive justice approach, such as the criminal justice system, schools, and workplaces [167, 180]. In recent years, there

has been growing interest within the fields of CSCW and HCI to adopt a survivor-centered approach to addressing online harm [157, 145, 121, 65, 44]. Researchers have also applied alternative justice models to address issues of online harm [145, 74]. My research joins these lines of work and explores how restorative justice can empower survivors in the online context.

However, applying restorative justice in an online context is not straightforward. For example, while offline restorative justice prevents harm and holds offenders accountable through a sense of community, the unique characteristics of online communities — such as anonymity [101], lack of social cues [45], and weak social ties [69]—create new challenges for developing a sense of community. While introducing and applying restorative justice requires resources and labor, content moderation also faces challenges in those areas [63, 131]. In my dissertation research, I ask the question:

> *How can we empower online harm survivors to address their needs with a restorative justice approach?*

I answer the research question from three interrelated perspectives. First, I understand the needs of online harm survivors beyond what traditional content moderation mechanisms achieve:

> *RQ1. What do survivors need in addressing online harm?*

I apply interview and design methodology to empower survivors to pinpoint their needs beyond what the traditional content moderation approach offers. Drawing from the offline restorative justice practice and speculative design, I developed a sticky note design activity to assist survivors in articulating their needs. This led to the formation of a taxonomy that segments adolescent needs into three areas: the specific nature of the needs, the actions and stakeholders responsible for addressing them, and the timeframe for such interventions. My research indicates that content moderation might not fully address survivors' needs, such as making sense of the harm, offering emotional support, and preventing subsequent harm. Nonetheless, several online and offline stakeholders, including bystanders, families, friends, and schools, can collaboratively tackle the problem. This taxonomy is a deeper reflection on the interactions and prioritization of the varied needs survivors might have. Additionally, it promotes the mobilization of resources from various stakeholders to address online harm faced by youth, including online communities, social media platforms, parents, and educational institutions.

Next, informed by my understanding of survivors' needs through a restorative justice perspective, I further develop social-computing systems to assist survivors. I focus on a one critical and initial need of survivors — sensemaking:

> *RQ2. How do we meet online harm survivors' needs of sensemaking?*

Sensemaking is a pivotal initial phase in addressing harm. It acts as a bridge, connecting survivors to resources and tools tailored for their needs. I designed and built SnuggleSense, an online platform that empowers survivors to make sense of the harm they have experienced and formulate a plan of action, especially in scenarios where there is a lack of immediate support or reluctance to seek assistance due to concerns of additional harm. While SnuggleSense primarily provides a reflective space for individuals, it also integrates community elements by presenting insights and suggestions based on the collective experiences of other survivors. Empirical results highlight that SnuggleSense aids in more effective sensemaking and empowerment for survivors compared to confronting harm without structured guidance. Our data indicates that SnuggleSense enhances survivors' sensemaking using a two-pronged community-based approach. This approach includes the survivors' existing social circles and an expanded network of survivors with analogous experiences brought together through SnuggleSense.

Enhancing the survivors' sensemaking process necessitates the provision of resources and tools, as well as the creation of a nurturing environment that empowers survivors to make informed decisions. In the final project of my dissertation, I situate survivors in the current moderation landscape and interrogate a multi-stakeholder process of addressing harm. I use restorative justice as a lens to examine the experiences of survivors, perpetrators, and community moderators in the Overwatch Gaming Community to explore the opportunities and challenges for practically implementing restorative justice:

> *RQ3. What are the opportunities and challenges of implementing restorative justice in the current moderation landscape?*

Through in-depth interviews with survivors, perpetrators, and community moderators, I identified challenges in integrating restorative justice into online platforms. These challenges stem from structural, cultural, and resource constraints, including the labor and resources required for education and deployment, the dominant punitive perception of justice, and the intricacies of asserting accountability online. To address these challenges, I proposed a phased approach, starting from adapting existing moderation practices to incorporate restorative justice values, such as centering impact on survivors in moderation explanations. As more stakeholders become involved and resources become available, a more comprehensive implementation becomes feasible. In this setting, I consider the function of restorative justice within the larger scheme of online moderation and stress the importance for platforms to re-evaluate their values. Additionally, I explore the ramifications when restorative justice endeavors fall short of their objectives.

The following chapters are laid out as follows. In chapter 2, I described the related work to my research and give some key concept definitions. In chapter 3, I reflect on my values as a researcher and how my values shape my approach to the research question. In chapter 4, 5 and 6, I detailed my research that answered the three research questions respectively. In chapter 7, I discuss the implication of my research and propose the future directions of my dissertation.

Chapter 2

Related Work

2.1 Online Harm and Content Moderation

Online harm

Online harm can refer to a myriad of toxic behaviors, such as public shaming [135], trolling [34], and invasion of privacy [105]. In this research, the harm cases we study can be characterized as interpersonal violence (compared to self-directed or collective violence), which is the harm that happens in the interaction between two or several individuals [98].

Online harm can impose severe consequences for the mental health and even the physical safety of its survivors [36]. Prior work has identified survivors' strategies of dealing with harm, including using mute or block functionality [83, 25], subscribing to bulk blocking mechanisms like Twitter blocklists [82] and collective story-telling [44]. Researchers have also sought to understand the perspectives and practices of perpetrators who perpetrate online harm as well as how they respond to current moderation practices [127, 80, 81].

Traditional approach of addressing online harm: content moderation

Online platforms currently address harm through content moderation, which usually involves punitive measures such as removal of content, muting, or bans [62, 131]. Online platforms' moderators can be volunteers who are users of the platform or commercial content moderators hired by social media companies [151, 131]. In recent years, social media platforms have also begun using automated, AI-based tools such as bots to help enact moderation [18, 91]. Current online moderation widely applies the *graduated sanction model*: the moderation of offending behaviors begins with persuasion and proceeds to more forceful measures [92, 125]. For example, an offender may receive a warning for their first offense, then a temporary ban for the next one, and finally a permanent ban.

As the most widely applied model of addressing online harm, content moderation faces many implementation challenges and is criticized for its limits in building healthy communi-

ties. One key challenge in implementing moderation is the sheer amount of labor required: as social media platforms grow, moderators must address increasing numbers of harm cases. Another challenge is the emotional labor required to moderate potentially upsetting antisocial content [131, 46, 178]. At the same time, moderation is not always efficient and effective in addressing harm: content policies and their implementations have failed to sufficiently remove disturbing material like fake news (a colloquial term for false or misleading content presented as news) [4], alt-right trolls [123, 134] and revenge porn [37, 168]. Researchers have examined ways to improve the current moderation processes, for example, through setting positive examples and social norms [150, 29], providing explanations for post removals [81, 79], and placing warning labels in front of inappropriate content [120].

Alternative approach of online governance and addressing online harm

In recent years, researchers have begun asking questions about the role that social media platforms play in the realization of important values related to public discourse such as freedom of expression, transparency, protection from discrimination, and personal security and dignity [160, 64, 70, 159, 161, 122]. Many scholars have offered ethical frameworks to guide platforms in their content moderation efforts. Perhaps the most prominent framework for ethical conduct in social media moderation are the Santa Clara Principles [139] that outline "minimum levels of transparency and accountability" and propose three principles for providing meaningful due process—publishing data about removed posts; notifying users affected by moderation; and granting moderated users a right to appeal. Shannon Bowen applies Kantian deontology to social media content management decisions, advocating that public relations practitioners rely on universal principles such as dignity, fairness, honesty, transparency, and respect when communicating via social media [23]. Carroll argues that social media sites are publicly traded companies and to be considered a good corporate citizen, they must fulfill their economic, legal, ethical, and philanthropic responsibilities [27]. Johnson [84] offers two ethical frameworks to guide platforms in governing speech: one grounded in an ethical concern for promoting free speech and fostering individual participation; and the other arguing that platforms owe users, whose content they commodify, a duty of protection from harms. Clearly, there exists a variety of frameworks, originally developed in offline settings, that usefully inform the ethics of content moderation [26]. Yet, the restorative justice framework is an especially informative framework for addressing online harm because of its focus on individuals and relationships rather than content; it therefore forms the focus of my research.

An emerging line of research that examines the values of a diverse range of stakeholders of online platforms to inform governance practices. For example, Schoenebeck et al. inquired about a variety of victims' preferences for moderation measures based on justice frameworks such as restorative justice, economic justice and racial justice [145]. Marwick developed an explanatory model of networked harassment by interviewing people who had experienced

harassment and Trust & Safety workers at social platforms [109]. Jhaver et al. conceptualized the distinctions between controversial speech and online harassment by talking to both victims and alleged perpetrators of harassment [80, 82]. Supporting such inquiries, Helberger et al. argues that the realization of public values should be a collective effort of platforms, users, and public institutions [70].

In particular, a growing body of research has embraced a survivor-centered approach to comprehending the experiences and needs of individuals facing online harm [157, 145, 121, 65, 44]. Studies have revealed that survivors possess needs that content moderation alone cannot adequately address, including the need to make sense of their experiences, receive emotional support, and contribute to the transformation of the online environment [181, 165]. There exists an urgent requirement to provide guidance, support, and resources to assist survivors in fulfilling these needs [164, 181].Researchers have studied survivors' experiences and challenges in addressing harm, exploring both collective sensemaking and empowerment [121, 19], as well as individualized endeavors and needs [76, 145, 20].

Researchers have studied how social-computing platforms can support people who experience harm. These studies often focus on people who experience harm in an offline context, such as pregnancy loss [11], depression [7], or the COVID-19 pandemic [104]. Most of these communities are on social media platforms and thus adopt the platform's content moderation to regulate the content [104, 7, 44]. Despite the scale of online harm and the unique challenges online survivors face, there are rarely communities that focus on the experiences of online harm survivors. In recent years, researchers have developed platforms or tools specifically designed to support online harm survivors in addressing their needs. Some tools try to provide social support by asking survivors' friends to review their messages [108] or by allowing survivors to share their experiences and receive advice and help for reporting online harassers [19]. Others have built tools to help online harm survivors document evidence of harassment [158, 65].

2.2 Restorative Justice

Restorative justice principles

As the culturally dominant way that we deal with harm, punitive justice is often most familiar to us. Therefore, to explain restorative justice, I will first contrast it with punitive justice. In Western cultures, the dominant model for justice when harm occurs is punishing the offender [162]. The punitive justice model holds that harm is a violation of rules and offenders should suffer in proportion to their offense [57]. The central focus of this model is on punishing and excluding the offender. However, the victim's concerns about the effects of the offense are rarely taken into account. Further, this model does not help offenders become aware of the negative impact they cause and take accountability to repair the harm [162]. Analyzing how harm is addressed in early MUD (multi-user dungeon) communities, Elizabeth Reid observed that "Punishment on MUDs often shows a return to the medieval.

While penal systems in the western nations...have ceased to concentrate upon the body of the condemned as the site for punishment, and have instead turned to 'humane' incarceration and social rehabilitation, the exercise of authority on MUDs has revived the old practices of public shaming and torture" [130]. Two decades later, the public spectacle of punishment is still prevalent on current digital platforms.

Restorative justice provides an alternative way to address harm. It argues that harm is a violation of people and relationships rather than merely a violation of rules [167]. It puts victims at the center of the process and seeks to repair the harm they suffer due to the offense. Restorative justice has three major principles [114]: (1) provide support and healing for victims, (2) help offenders realize consequences of their wrong-doing and repair the harm, and (3) encourage communities to provide support for both victims and offenders and to collectively heal. There can be multiple levels of communities primarily affected by harm, including the local community where the harm occurs or the broader society [114]. Our study locates victims and offenders in the Overwatch gaming community.

Restorative justice practices

Restorative justice has been successfully applied in a myriad of settings, such as the criminal justice system, schools and workplaces [167, 180]. When formalized within an organization, restorative justice processes can be embedded within a punitive justice system to use on selected types of harm cases [167, 180]. Those who do not want to proceed using this approach or cannot reach consensus during the restorative justice process are directed to the punitive justice system [13]. A widely used practice in restorative justice is the *victim-offender conference* [184]. Here, victims and offenders sit together with a restorative justice facilitator to discuss three core questions: (1) what has happened? (2) who has been affected and how? (3) what is needed to repair the harm? The facilitator mediates this process to ensure that victims and offenders have equal footing and helps move the parties towards reaching consensus. Other forms of restorative justice meetings, such as family-group conferences, include other community members such as family and friends of victims and offenders [184, 13]. Restorative justice practices embed values and principles of restorative justice to meet the needs of all parties involved, including the victim, offender and community members. In this paper, we use victim-offender conference as an exemplar to inquire about participants' preference for online restorative justice practices.

Restorative justice practitioners emphasize the importance of preparation in advance of the victim-offender conference. A restorative justice facilitator first meets separately with the offender and the victim before the conference in a process called a *pre-conference* [184]. During these meetings, the facilitator introduces the restorative justice framework to the victim and the offender and asks them the same set of questions that will be asked during the conference. After these meetings, the victim-offender conference happens only when both parties agree to meet voluntarily to repair the harm. Additionally, the facilitator acts as the gatekeeper to determine whether victims and offenders can meet to reach a desired outcome without causing more harm. In both pre- and victim-offender conferencing, the

facilitator does not make decisions for victims or offenders about how to address the harm, but guides them to reflect on the harm through the restorative justice framework [22]. For this research, the first author received training and acted as a facilitator in pre-conferencing with the participants to ask about their attitudes towards engaging in a restorative justice process to address online harm.

Restorative Justice outcomes

Restorative justice does not encourage punishment as the desired outcome. Instead, it focuses on the obligations to repair harm and heal those who have been hurt [184]. Possible outcomes of a restorative justice conference include an apology from the offender, or an action plan that the offender will carry forward, e.g., doing community service or attending an anger management course [40]. Restorative justice framework acknowledges that it can be difficult, or in some cases, even impossible to fully restore the situation or repair the damage. However, symbolic steps, including acknowledgement of impact and apology, can help victims heal and offenders learn and take accountability [128]. When restorative justice is embedded within a punitive system, the outcome of restorative justice can inform the punishment decision in some cases [167]. For example, when victims and offenders can reach a consensual outcome in a restorative justice process, offenders may receive reduction or exemption from a punitive process. In other cases, a restorative justice process may run in parallel and have little influence on the punitive process [167].

Here, an example may be instructive. A high school in Minnesota was dealing with problems of drug and alcohol abuse. The school held a conference that gathered the offender (a student who had used drugs on school grounds), affected students, and members of the faculty and staff. The offender first shared her story and the reasons for her actions. She also took the opportunity to seek forgiveness. Other members of the conference then expressed how they were affected by the offender's behavior and jointly discussed solutions. The outcome was that the offender became aware of the effects of her actions and agreed to go through periodic checks to monitor her continued sobriety [89].

While practicing restorative justice benefited the affected parties in the case above, does it always succeed? Latimer et al. conducted an empirical analysis of existing literature on the effectiveness of restorative justice and found that restorative justice programs successfully reduced offender recidivism and increased victims' satisfaction with the process and the outcome [102]. However, these positive findings were tempered by the self-selection bias inherent in restorative justice practices. Since it is a voluntary process, those who choose it may benefit more than others [102]. Restorative justice also requires commitment at the administrative level [59] and time and labor for the parties involved [184].

A restorative justice approach to addressing online harm

There is a growing interest in the research community to implement restorative justice values to address online harm. Blackwell et al. first introduced restorative justice in the content

moderation context [20]. Schoenebeck et al. conducted a large-scale survey study and showed that restorative approaches such as using apologies to mitigate harm were strongly supported by participants [145]. West proposed that education may be more effective than punishment for content moderation at scale [176]. Kou argued that permanent bans produce stereotypes of the most toxic community members and such bans prove to be ineffective over a long term. Kou recommended that instead of dispensing bans, online communities use a restorative justice lens to re-contextualize toxicity and reform members into becoming well-behaved contributors [95]. Hasinoff and Schneider examined the tension between the online platforms' desire for scalability and the restorative and transformative justice ideal of local contextualization. They argued that subsidiarity, the principle that local social units should have meaningful autonomy within larger systems, may address this tension [68].

Chapter 3

Positionality Statement

As a researcher working in the sensitive space of understanding online harm and exploring interventions to address it, I reflect on my position regarding this topic. I live in cultures where punitive justice is the dominant approach to addressing harm, and I have actively sought to understand the limitations and benefits of restorative justice through both research and personal practice. I am deeply concerned that online harm remains a persistent social issue, disproportionately affecting marginalized and vulnerable communities [48]. My prior research on misinformation and online harassment has highlighted the limitations of applying punitive approaches in existing moderation mechanisms, driving my interest in exploring alternative justice frameworks. I am encouraged by the growing interest in applying alternative justice theories [15, 75, 93, 116, 175], including racial justice [145] and transformative justice [38], to reduce harm for survivors, perpetrators, and communities. I am particularly inspired by the success of restorative justice in addressing offline harm and its potential to provide agency and care for vulnerable groups often overlooked or further harmed by punitive justice models.

While I embrace the values of restorative justice and its potential, I do not advocate for it uncritically in my work. I recognize the importance of different justice models. Traditional content moderation, as a punitive approach, has proven effective in halting ongoing harm and reducing re-offense [62, 79]. Addressing systemic issues like sexism and discrimination requires a transformative approach to tackle underlying structural problems [51]. I acknowledge the limitations of restorative justice in addressing deep-rooted, systemic cultural and social issues such as racism and sexism (see transformative justice [38]). In addition, restorative justice is a voluntary-based approach, and applying it universally without consent risks causing harm to both survivors and the community [184]. Rather than championing one specific approach, my goal is to empower survivors by drawing from alternative methods of addressing harm that are not traditionally adopted, thereby expanding the toolkit available to online harm survivors in addressing the harm they experience.

Chapter 4

RQ1. What do survivors need in addressing online harm?

4.1 Introduction

Online harm such as harassment is prevalent on social media platforms. According to Pew Research Center, social media is by far the most common online venue for harassment in the United States — 75% targets of online abuse, which equals 31% Americans say their most recent harm experience was on social media [170]. Social media platforms tend to address these harms through the framework of content moderation: the review and removal of content that violates the platform's rules, and banning of repeat offenders [62, 131]. Though research has found some impact of content moderation in reducing offenders and offending behaviors [79, 82], the framework leaves out victims' experiences and needs for addressing harm [145]. Research has found that the current form of content moderation leaves victims out of the decision making process [145] and fails to adapt to their individual experiences [19].

In recent years, the HCI and CSCW communities have explored a victim-centered perspective to address online harm. Researchers have examined victims' strategies for dealing with harm [25, 169], engaged victims in designing interventions to address online harm [9], and studied their notions of justice [20, 145]. Our research builds on this line of work and is inspired by restorative justice – a victim-centered justice approach – to understand people's needs for addressing online harm. Content moderation follows a punitive justice approach, where it responds to harm by centering the offending party and regulating their offending behavior through punishment. Restorative justice, on the other hand, centers victims' experiences and desired outcomes. Through communicating with victims, a restorative justice process aims to support victims to reflect on their needs for addressing harm, and also engaging offenders and community members to help victims meet those needs [184].

We focus our investigation on adolescents (10-20 years old) [140], which are a particularly vulnerable group for a variety of harms in the online space. The vast majority of teens (90%) in the United States believe online harassment is a problem that affects people their

age, and they mostly think teachers, social media companies and politicians are failing at addressing this issue [8]. Restorative justice has been successfully applied to address harm among adolescents in schools [72]. In recent years, researchers have seen the potential for applying restorative justice principles and practices in the online space [20, 145, 74]. We follow this line of work and explore how restorative justice helps us understand adolescents' needs in addressing online harm. Human needs and motivations are the driving force of behavior [113]. Without understanding needs and motivations, we may presuppose why certain actions (e.g., moderation) are important for addressing harm, but not understand why they are important for victims [126].

We examine adolescents' needs for addressing online harm from three interrelated perspectives: *what* needs they identify, *how* they believe their needs can be met (and by whom), and *when* they believe different needs should be met. Before we can develop specific recommendations to address the needs, we must first understand the types of needs that adolescents identify from their own experiences. For example, adolescents may identify needs for themselves, as well as needs for their online communities. Our goal in the first research question is to better identify specific, major types of needs that come from adolescents experiences with online harm:

> RQ1. What types of needs do adolescents identify for addressing online harm?

Next, we examine *how* adolescents hope to achieve their needs, including the relevant stakeholders and the actions they perform in order to meet those needs. When harm happens, victims suffer from a lack of agency and require actions from relevant stakeholders (e.g., moderators, bystanders) to collectively address the harm [82, 163].

> RQ2. How do adolescents want to achieve different needs in addressing online harm? What specific actions can help adolescents address their needs, and by whom?

Adolescents may have multiple needs for addressing harm, which requires more than one action from a single stakeholder. Needs are not necessarily independent from one another, and some needs may have more immediacy than others. For example, social and self-esteem needs can become more important once fundamental needs of safety and security are achieved [110]. In our research, we aim to understand the immediacy of needs in addressing online harm:

> RQ3. When do adolescents hope to achieve different needs when online harms occur?

To address our research questions, we conducted interviews and design activities with 28 participants who experienced online harm during adolescence. In the interviews, participants complete a series of task on the online whiteboard to identify and reflect on stakeholders, actions, needs, and the timeline to achieve those needs. We found five major needs for

addressing online harm from participants: sensemaking, support and validation, safety, retribution, and transformation. Participants identified both online and offline stakeholders that may address needs, including moderators, offenders, family and school, and proposed actions for address needs both in the short and long term.

In this paper, we examine what it would take for social media spaces to realize important social values such as supporting the safety and growth of adolescents, instead of the bare minimum of banning some types of offending content. Our findings shed light on how we may expand our understanding of victims' needs both spatially and temporally. We argue that online platforms can implement approaches beyond content moderation and can collaborate with other stakeholders to support victims both in the short and long term. In particular, we see potential for applying restorative justice approaches in addressing online harm for adolescents, such as by helping offenders realize their wrong-doing or utilizing the support of communities that victims are a part of. The design task that we created, contributes an innovative method for victims to reflect on their needs in addressing online harm. Finally, our research builds on and extends recent work that center victims' perspectives in addressing online harm [19, 142] and examines alternative justice models in online governance [145, 20].

4.2 Method

Our research aims to understand adolescents' needs for addressing online harm, including what those needs are, how to meet those needs, and when. While asking people what they need may seem like a straightforward task at first, prior work in the restorative justice literature and our own preliminary research showed that it is challenging for victims to know and express what needs they have [22]. This is particularly the case when when those needs were not met when the harm happened, or when meeting those needs seems impossible given available resources from the online platforms or other relevant stakeholders.

Victims of harm need to go through a process of sensemaking to understand the harm, its effects on them, and to decide what they need to heal from the harm [184]. In restorative justice practices, this is often done through a pre-conferencing session with a facilitator who support the victim and helps them figure out what they need [128]. For this research, we hope to design a task to support the process of sense-making and enable participants to tell us the whole range of their needs – even those that could not be immediately met given current constraints and resource limitations. In this section, we first described the process of designing the task. Next, we presented the task procedure. We then explained our recruiting and interview process, and finally ended with a description of our data analysis method.

Designing the task

Designing the need-finding questions to understand types of needs and actions

The goal of our research is understanding adolescents' needs when they are harmed online from three levels: (1) what their needs are, (2) the actions to meet those needs, and (3)

the timing to meet those needs. It is challenging for people to know what their needs are and how to express them. Thus, our goal is to design *need-finding questions* to help support participants' sensemaking process.

Weick argues that sensemaking is retrospective. People first come up with or perform actions, then provide explanations for their actions [174]. Thus, we focus on actions in the need-finding questions, and then ask participants to explain their actions. Through the explanation, participants can identify their needs behind those actions. The process enables us to answer RQ1 and RQ2 together; by understanding what peoples' needs are, as well as the actions that can meet those needs.

We aim to design need-finding questions that cover all the categories of actions participants may identify. We started our research design process by looking at how victims talk about needs and actions for addressing harm in the restorative justice literature. In Zehr's foundational work on restorative justice, he proposes four categories of needs that victims commonly have: the need for information, the need for truth-telling, the need for empowerment, and the need for restitution or vindication [184]. Zehr also describes example actions that can address those needs, for example, offenders' acknowledgement can meet victims' need for restitution, and understanding why the harm happened can meet victims' need for information. We rely heavily on this work in designing our research method.

First, the research team brainstormed potential actions based on the examples provided by Zehr. We then categorized the actions through pilot testing with 15 participants who we selected through convenience sampling [133]. We asked pilot participants with experiences of online harm to select from those actions, and come up with additional needs they might have. For pilot participants who hadn't experienced online harm, we asked them to group the actions in a card sorting activity [153]. We asked pilot participants to think out loud to understand their thought process. This process led to five questions which cover most actions our pilot participants mentioned: *(1) what information do you need from [the stakeholder]? (2) what do you want to share with [the stakeholder]? (3) what acknowledgement / understanding do you want from [the stakeholder]? (4) what actions is needed from [the stakeholder] to repair the harm? (5) what change do you want [the stakeholder] to do in the future?*

Designing a timeline to envision the story

While the need-finding questions help us answer RQ1 and RQ2, we were also interested in the temporal aspects of addressing online harm and envisioned a story line of addressing harm for RQ3. The process of participants walking through their own storylines provides more chances for them to reflect on their different needs and actions, as well as their order when a harm occurs. Inspired by previous research in speculative design [179], we decided to design and facilitate a reflection process to achieve this goal.

Our design task borrows from the Timelines speculative design activity proposed by Wong and Nguyen [179]. Timelines is designed to help participants reflect on their values and ethics around a technology. Participants complete the Timeline activity with sticky

notes and a whiteboard. There are four steps in the timeline activity: (1) participants decide on an artifact (e.g., a technology) as the topic of discussion, (2) identify stakeholders around the artifact, (3) create potential news headlines and stories related to the artifact, and (4) organize the news headlines and stories on multiple timelines to create stories of events related to the artifact. Overall, through a visual board, the Timelines activity helps "the creation of an imagined world that can lead participants to critical refection" [179]. In borrowing from the Timelines activity, our goal is to help participants picture a storyline for addressing harm, while reflecting on their values and desired outcomes in the process. While our work is not entirely speculative, we encourage participants to think beyond perceived constraints while building on their own experiences.

Task procedure

Our study consists of four main stages:

1. Participants decide on a harm case from their adolescence they'd like to talk about.
2. Participants identify stakeholders relevant to the harm case.
3. Participants generate actions the stakeholders might perform with five need-finding questions. Participants reflect on their needs through identified stakeholders and actions in stage 2 and 3.
4. Participants map those actions spatially to illustrate their preferred timeline for addressing the harm.

Figure 4.1 provides an example of the interface where participants complete the task. In the following sections, we describe each stage in more detail.

Stage 1: Participants choose a harm case from their adolescence they'd like to talk about.

In the first of four stages, the researcher asks participants to share a harm case they want to talk about. In pilot studies, we shared a hypothetical harm scenario with participants and asked them to imagine themselves in that situation and share their needs. However, participants found it hard to empathize with the hypothetical scenario. Therefore, we chose to use participants' lived experiences. While relying on each person's own experiences made it more difficult to control for the types of harm in our study, their personal experiences contain concrete details (e.g., what the offender said to them, their relationship with different stakeholders) that are important to determining their needs.

After participants select a case, we ask a series of questions about the case (e.g., when and where it happened, if and how they addressed the harm, and their feelings at that time). The purpose is twofold: first, in later stages of the task, we provide participants flexibility

Table 4.1: Participant demographics and experiences of online harm

	Age	Gender[1]	Race / ethnicity[1]	When harm occurred	Online Platform	Offline Site	Number of offender(s)	Relationship	Description of harm[2]
P1	19	Female	Asian	Middle school	Instagram, Snapchat	School in U.S	1	Friends	Racist comments, public shaming, physical harm
P2	20	Male	White	17-18	Twitter	N/A	1	Stranger	Physical threat
P3	20	Female	Asian	15	Discord	N/A	1	Friends	Sexual harassment, non-consensual image sharing
P4	19	Female	Asian	16	Instagram	N/A	1	Schoolmate	Racial discrimination, public shaming
P5	20	Male	Asian	20	League of Legends	N/A	1	Stranger	Offensive name-calling
P6	20	Female	Asian	Middle school	Instagram post	School in U.S	3 to 4	Friend	Public shaming, non-consensual image sharing
P7	20	Female	Asian	High school	Instagram	N/A	1	Classmate	Body shaming
P8	18	Female	Indigenous	18	Twitter	N/A	3 to 4	Strangers	Racist comments, offensive name-calling
P9	19	Female	Asian	First year of high school	Instagram	N/A	1	Classmate	Body shaming
P10	19	Male	Asian	End of middle school	WhatsApp	School in India	10 to 15	Classmate	Public shaming, physical threats, physical harm
P11	19	Male	Hispanic white	18	Instagram	N/A	1	Friend	Making fake profile of me
P12	19	Female	Asian	19	Tiktok	N/A	Multiple	Strangers	Racist comments
P13	19	Male	Asian	19	Grindr, tinder	N/A	Multiple	Strangers	Financial fraud with fake account
P14	20	Male	Asian	20	Reddit	N/A	Multiple	Strangers	Trolling, harassment
P15	20	Male	Asian	High school	Instagram, Facebook Messenger	School in U.S	Multiple	Friends	Racist comments, making jokes of my disability
P16	19	Female	Asian	High school	Instagram, Snapchat	N/A	Multiple	Friends	Non-consensual image sharing
P17	20	Male	Asian	First year of high school	Instagram	School in India	1	Friend	Making fake account for public shaming
P18	20	Female	Asian	High school	Instagram, Snapchat	School in China	Multiple	Classmate, strangers	Racist comments
P19	20	Female	Asian	13	Tumblr, Insagram	N/A	Multiple	Strangers	Offensive comments of my arts
P20	19	Female	Asian	High school	Instagram	N/A	Multiple	Strangers	Racist comments
P21	19	Female	Asian	18	Slack, email	N/A	1	Stranger	Non-consensual image sharing
P22	20	Female	Asian	18	Instagram, Twitter	N/A	Multiple	Strangers	Racist comments
P23	20	Female	Black	Elemetary to high school	ASKfm	N/A	Multiple	Schoolmate	Sexual harassment
P24	18	Male	Asian	High school	Facebook	School in India	Multiple	Classmate	Body shaming
P25	19	Female	Asian	High school	Weibo	School in China	Multiple	Classmate, strangers	Trolling, harassment
P26	20	Male	Asian	High school	Twitter	School in U.S	Multiple	Schoolmate	Racist comments, offensive name-calling
P27	19	Female	Black	10th grade	Instagram	School in U.S	Multiple	Schoolmate	Public shaming, harassment
P28	20	Female	Asian	20	Instagram	N/A	1	Stranger	Racist comments

[1, 2]Participants' gender and race/ethnicity are self-identified. The majority of participants are Asian, yet they come from diverse cultural background, including Afghan, Chinese, Filipino, and Indian. [3]Here, we did not follow a strict definition of types of harm but rather stay close to participant's description and categorization of their experiences.

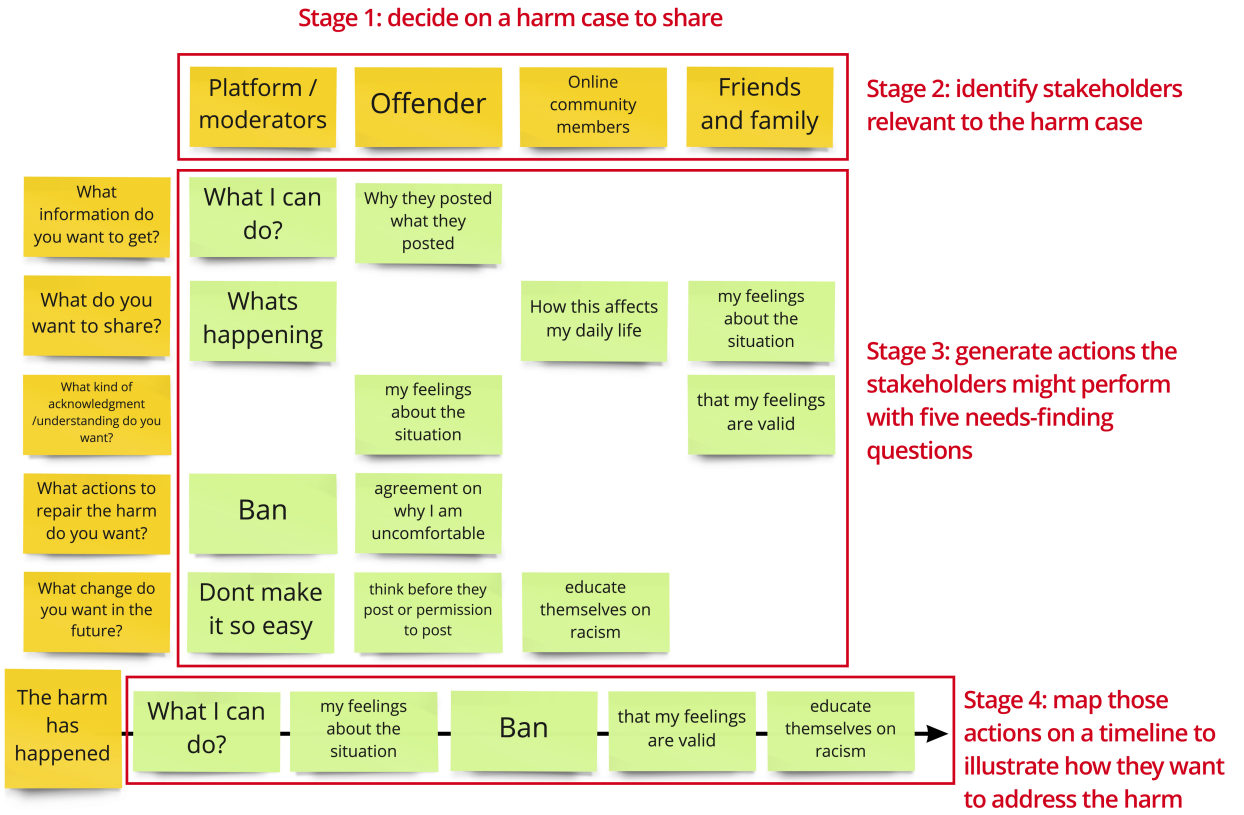


Figure 4.1: An example of the interface where participants complete the four-stage task. The sections marked by red frames are what that participants need to complete in each stage of the study. To protect participants' privacy, the data in the red frames comes from multiple participants.

in expressing the needs based on their unique experiences. Talking about the harm they experienced helps participants recall what has happened in detail, which enables them to reflect on their needs thoroughly. Second, information about the harm case provides context for our interpretation and understanding of their needs in data analysis.

Stage 2: Participants identify stakeholders relevant to the harm case.

In the second stage, participants identify the stakeholders relevant to the harm. We asked two questions to facilitate their selection of stakeholders: *(1) Who is responsible to help you address the harm? (2) Regardless of responsibility, who can support or help you to address the harm?* Because some participants weren't sure how to answer this question in pilot studies, we provided them with some example stakeholders as starting points, including offenders,

family and friends, online community members, and platform/moderators. Participants also have the option of adding additional stakeholders either in this stage or in later stages.

Stage 3: Participants generate actions the stakeholders might perform with the five need-finding questions.

In the third stage, we used the stakeholders they identified and five need-finding questions discussed in section 3.1.1 to form a table, and asked participants to answer each question with respect to each stakeholder group in the table (see Figure 4.1). Thus, the process allows participants to identify the *actions* required from stakeholders for addressing the harm.

In stages 2 and 3, we asked the participants to think out loud and explain their rationale for selecting/writing a note. This allows us to understand the type of *needs or motivations* of choosing certain actions or stakeholders.

Stage 4: Participants map those actions on a timeline to illustrate how they want to address the harm.

In the final stage, participants rearranged the actions they had just created spatially to reflect on an ideal timeline to address the harm. Since participants have reflected on the needs behind the actions, the series of actions on the timeline also represents the sequence of needs. We also encouraged them to create new notes to complete the timeline. In the process, the researchers asked participants to think out loud and explain their reasoning for the order of notes.

Recruitment and Interviews

We recruited 28 students from a University on the West Coast of the United States. We used the university's internal recruiting platform to reach potential participants from a pool of students who agreed to be contacted about paid social research opportunities. We focused our recruiting message to late adolescent students between 18-20 years old, and indicated that we were looking for participants who have experienced online harm on social media during their adolescence. We also provided some examples of online harm (e.g., offensive name-calling, public shaming, stalking, harassment, physical threats) to help them reflect on potentially relevant experiences.

We show participants' demographic information and the information about their online harm experiences in table 6.1. Participants reported a wide range of harm experiences including non-consensual image sharing, body-shaming, sexual harassment and physical threat. For many participants, the harm cases had an offline part in school, or happened online with their classmates or schoolmates. Due to the restriction of IRB, the participants we interviewed are in the late adolescence group, but the harm cases they shared happened from early to late adolescence.

We conducted all the interviews between July and August 2021. The interviews were within one hour and in the form of video or voice call on Zoom (www.zoom.us). We used an online whiteboard tool, Miro (http://miro.com), to facilitate the task in remote sessions. Participants received a $25 gift card as compensation. The study is approved by the institutional review board.

Data Analysis

We transcribed the interview recordings with an online transcription service (www.rev.com). We exported the data on the online whiteboards into an excel sheet, and also referred back to the original board for spatial information during analysis. We analyzed the interview transcript and data from online whiteboards together.

We conducted the data analysis in an iterative process. We applied interpretative qualitative coding to the data [115]. We began with initial coding, where we applied short phrases as codes [136]. The first round of coding was done on a line by line basis so that the codes stayed close to the data. Some example codes include "need empathy" and "design automatic moderation tools." We then conducted focused coding by identifying themes that appeared repeatedly to form higher-level descriptions [136]. Examples of second-level codes include "prevention of harm" and "acknowledgement of responsibilities." Throughout the analysis, we not only paid attention to the needs victims have, but also the actions and stakeholders that were proposed to meet the needs. In analyzing the timelines data, we paid attention to the order and time span in which participants wanted to address those needs.

4.3 Findings: Stakeholders and Actions

Our first research question concerns the types of needs that adolescents identify for addressing online harm. Our second research question explores their preferred actions and stakeholders to address those needs. As we noted in methodology, it is challenging for participants to directly identify their needs. Thus, we first asked participants to identify their preferred actions and stakeholders for addressing harm, and then asked them to explain the needs behind the actions and stakeholders retrospectively. Since needs, actions and stakeholders are three interrelated concepts, we answer the two research questions together in this section. We presented our major findings in table 6.2. We found five major needs that participants frequently mentioned in our study: sensemaking, support and validation, retribution, safety and transformation. Next, we detail the actions and the relevant stakeholders related to those needs.

Need for sensemaking

Through the reflective task, all participants indicated that the offenders had done something wrong. However, some participants told us that they were not as certain about the

Table 4.2: The table presents participants' needs for addressing harm, actions to meet the needs, and the stakeholders to perform the actions from left to right. For example, to meet the need of sensemaking, participants hope to seek information on offenders' motives for conducting harm from their offenders.

Needs	**Actions**	**Stakeholders**				
		offenders	**family and friends**	**online platform**	**online community members**	**school**
Sensemaking	Seek information on offenders' motives for conducting harm	x				
	Seek advice on how to address the harm		x	x		x
Emotional support and validation	Emotional support		x		x	
	Show stance against harm			x	x	
	Acknowledge wrongdoing and issue apology	x				
Retribution	Content moderation			x		
	Report or call out offenders				x	
	In-school repercussion					x
Safety	Acknowledge wrongdoing and issue apology	x			x	
	Show stance against harm				x	
	Content moderation			x		
	In-school repercussion					x
Transformation	Improve design and moderation			x		
	Raise public awareness				x	x

wrongdoing immediately after the the incident had occurred. P3 reflected that her offender *"[repeatedly] sent me unwanted pictures then brushed it off as a mistake."* She was unsure about the situation: *"That were almost in a gray zone [. . .] I was not sure if that was just how he normally is and I'm just overreacting or is he actually making unwanted advances towards me."* Additionally, even when some participants were sure that they were harmed, they were less sure how to address the problem. To make sense of the situation, participants hope to understand why offenders did what they did, and get instructions or advice on how to deal with the harm.

Seeking information on offenders' motives for conducting harm

Participants hope to get information on why the harm happened, and whether they are actually responsible for the harm instead. Several participants mentioned that they want to get the offenders view, and to understand their actions. P28 explained that it would help her to *"understand that the problem is within themselves [offenders] and not with me."* P5 expressed that understanding offenders' motives helped him rationalize offenders' behavior:

"maybe they're having a really bad day then it makes more sense I feel for them to behave that way." Understanding where the offender is coming from is also a step towards addressing the problem. P6 said, *"I wanted to know why it happened, so we were able to talk about it."*

Seeking advice on how to address the harm

Many participants mentioned that they need advice from others in dealing with the harm. Some of them chose to resort to their family and friends. P18 believes that family and friends can give *"personalized advice"*. Other participants want to look for people who have expertise in addressing harm. P10 explained that mental health professionals can help with his emotional sufferings: *"they just know techniques that people can use to cope with any sort of suffering and how to alleviate it."* P6 described the needs for online platforms to help address some technical challenges: *"I didn't know how to protect myself. Even now I don't fully understand how to do things on Facebook, prevent people from tagging me in photos that I don't want to be tagged in or stuff like that."* P10 also identified the guidance counselor from school as someone who could help with harmful situations, particularly since many of the harms happened among classmates: *"They're used to dealing with school kids. They know how bullying happens. They know the triggers."*

Need for emotional support and validation

Almost all participants mentioned the need for getting emotional support and validation. Participants look for emotional support from family, friends and online supporting groups, while hoping platforms and online community members will acknowledge their responsibility in addressing the harm. In some cases, they hope the offenders can acknowledge their mistake and apologize.

Emotional support from family, friends and online users with similar experiences.

When a harm occurs, participants are charged with negative emotions. Participants explain that they need to vent their feelings and hope to get emotional support. Some resort to friends and family for those conversations: *"Because with friends and family, you can really open up about how you're feeling...whether you just want to rant, or you want advice, then they can either offer that support. Because they probably care more than most people" (P2).* Some participants also turn to other online users who have similar experiences. Here, participants gain support not only through sharing their own stories, but also by listening to others. P3 said, *"Support communities are a good place to share experiences without fear of being judged, and survivors tend to feel more solidarity when hearing about others' stories."*

Platforms and online community members can show their stance against harm

Some participants believe that online community members and online platforms can provide support and validation by standing in solidarity with the victims against online harm. P19 explained that she hopes online users realize that they are connected: *"It's important to know that we're all one big community that's sharing something...it's important to build each other up."* P20 expected online community members to express *"kind of a collective understanding that what was happening was wrong and there should be preventative action against it."* Several participants hoped that platforms could issue direct statements to show their stance towards the incident (or similar incidents). P7 provided an example: *"[I hope the platform can state that] 'these types of behaviors and comments are not appropriate in any setting."'* Participants explained that acknowledgement of online harm is one step towards addressing the issue: *"I guess to repair the harm, acknowledging that there is a problem is the first step" (P18).*

Offenders can acknowledge their wrong-doing and issue an apology

Some participants explained that they want the offenders to acknowledge the harm that they created. For example, P2 hoped the offender might understand that *"it was really hurtful. It was hateful. It was unnecessary."* Participants not only hope to share their feelings and frustrations, they often want an acknowledgement of wrongdoing from the offender. In particular, several participants explained that the offenders can (or should) apologize to them. P1 wants to tell her offender *"this is how you made me feel, and this is how interpreted the situation. please understand my side"*, and in return, the offender should *"[apologize] and clear the air."*

Need for retribution

Participants often explained that they hope offenders receive consequences for their negative behaviors: *"you'd want to send a message [to offenders] that hate will not be tolerated...that actions have consequences" (P14).* Some participants specifically mentioned that they want offenders to receive punishment as consequence. After being harmed repeatedly, P27 admitted that she hoped the offender would suffer in return: *"I used to be really angry and I just wanted bad things to happen to them...they shouldn't get away with it."* While some participants expressed a need for punishment, restorative justice explicitly seeks to create alternatives to punitive justice [184]. We will discuss this conflict, as well as the difference between accountability and punishment, in the discussion section.

Participants described different authority figures who might administer retributive actions. Some believe that online platforms can use moderation (e.g., bans) to hold offenders accountable. Participants explained that online community members could help report offenders to moderators, or call out offenders at the time of the offense. Since many participants

receive harm from their classmates or schoolmates, some believe the school administrators should hold their students accountable.

Platforms: issue punitive moderation decisions to the offender

Some participants believe that platform moderation (e.g., banning, muting) is a form of punishment to offenders or a way to hold them accountable. P23 explained her rationale: *"I think that your presence on social media is a privilege that can be taken away...if you don't follow the rules or the guidelines of the platform."* P18 thinks that banning is a form of denial to offenders: *"suspension of a rude account isn't really a big thing, but...personally, I think that would make me feel better that the people who were rude to me, someone is telling them that what you did [is] wrong."*

Bystanders: Reporting offenders or calling out offenders for their actions

Participants also mentioned online community members, in particular, bystanders' role in holding offenders accountable. Some participants believe that bystanders should report the offenders to the platform. P6 hopes bystanders know that *"It is important and very simple online to report things that you see that are harmful to others."* In addition to officially reporting harm, some participants expressed that bystanders could call out the offending behavior when it occurs. P28 described it as *"a form of positive online peer pressure"*, while P23 phrased it as *"public backlash."*

School Administration: Punishing students for their online behavior

Some participants believe that the school should hold their students accountable for their online behavior. P24 explained, *"It's your [the school's] duty to ensure that the students of your school behave in a good way...and then train them to be good citizens. So that's why I feel like, even if it's online, they [the responsibilities] still go to your school."* Participants believe that the school should give students academic repercussions for their online behavior. P27 described it as *"getting detention or something showing up on their records to show that they have poor behavior."* P18 stated that the school should *"Talk to their [offenders'] parents, maybe even suspend them."*

Need for safety

We find that sometimes participants need to deal with an ongoing harm. Even when participants think that a harm has stopped for the moment, they are often unsure if the harm will resume in the future. In such situations, one's safety from continuing harm is a priority. Previously, we explained how a variety of actions can meet participants' need for support and validation, as well as a desire to get retribution from offenders. Importantly, we find that these *same* actions can serve another purpose – to stop the continuation of harm and help individuals feel safe.

Participants hope that acknowledgement can stop the harm.

Earlier we described how participants need emotional support and validation after the experience of harm. They get comfort when online community members show their stance against online harm, or when offenders acknowledge their wrong-doing and apologize. Some participants believe those actions also stop the continuation of harm. P19 believes that online community members' stance against harm can reduce offending behavior: *"If it's publicly announced that, 'Oh, this is not the behavior that we're going to tolerate,' I feel like people would be more ashamed to act out like that."* Some participants also expect their conversation with the offender can prevent continuous harm from them: *" [If offender] apologize and clear the air between us, and then any acts of harm should stop because we are done with it" (P1).*

Retributive actions to stop the harm

Some participants hope that retributive actions will teach the offender a lesson, while also stopping the harm. We find that participants often put the onus on platforms to enact some type of retribution which will also stop the harm: *"If they [offenders] are not willing to change, it's kind of the responsibility of the platform to kick them off"(P19).* P24 explained why having the harmful post reported and removed can stop the escalation of harm: *"you want to ideally reduce the number of views that it gets and prevent it from growing even bigger."* In addition, some participants expect the school can intervene: *"the school should step in and say, 'stop taking people into this [the harm]"' (P6).*

Uncertainty about stopping the harm

Participants in our study do not always expect that having a conversation with offenders will force a change. P7 explained, *"The offender would always try to defend themselves I think, and not really address anything."* She believes that *"people don't need don't really change overnight."* P15 expressed reluctance to face the offender and worried if he will be disappointed by the response: *"There is the sort of fear that maybe they won't understand...if you don't get the response that you're ultimately looking for, then it can just be uncomfortable."*

Even when participants rely on others for help, they sometimes cannot specify the actions they want others to perform, or know whether those actions will effectively stop the harm. Individuals may know that they want something to stop, but they do not know who should actually take action. P6 expressed this type of frustration, *"I feel like someone should put a stop to that."* P10 put his need as a inquiry: *"I would want to know what they [moderators] can do to stop these kind of hateful messages from being spread, so that they could put a halt to the situation and at least [lower] its severity."* P6 was also unsure whether the school can hold their students accountable for online behaviors: *"ultimately they don't have the physical capability to make it stop...if those students were not willing to be compliant, I'm not really sure what actions you could expect the school to take."*

Need for transformation

Some participants believe that addressing online harm not only mean working on their individual cases, but also fundamentally change or transform the online environment. Participants suggest that it is important to bring more attention and resources to the issue of online harm. These individuals believe that online platforms should use more resources to improve their moderation procedures, and emphasize online environments that identify and stop harms from occurring. At a broader level, these participants indicate that it is important to raise public awareness of online harm.

Online platforms should improve design and moderation to address harm

Several participants indicated that the platforms should moderate content before a harm occurs. For example, several participants believe that platforms can improve their automatic detection mechanisms to filter out hateful comments before they reach online users. P7 thinks the current detection only works for explicit hateful words: *"Some comments they filter are usually only addressed for inappropriate things as in rated R things inappropriate."* He hopes platforms can *"filtering comments that are close to hate or bullying."* P18 also expressed that human moderators can *"keep an eye out for certain rude words or phrases."*

Participants explained that platforms can provide more information, tools and resources in response to people's individualized experiences. P11 wanted to correct a fake account someone made of him, and he hopes the platform can demonstrate *"the process of investigating and compiling a report on a duplicate social media account."* P10 argues that every harm is different: *"like every person's experience is nuanced and sometimes it really just doesn't fit in one category. In my case I actually knew the person, so it'd be nice for you to have more of a place to talk. "* He told us that only human moderators can provide customized responses: *"unless and until a human being looks at it and figures out what's going on, I don't think that's very accurate."*

Raising public awareness about harm can improve online environments

Many participant believe that educating the public about the importance of online harms and how to deal with them can fundamentally improve online environments. P18 talked about how she resorted to her family for help when online harm happened but wasn't given enough attention. She thinks that online harm was a new concept to her elderly family members, which she wanted them to understand: *"For them, bullying was something in person and rude comments on Instagram isn't even considered bullying for a lot of people."* Other participants explained that they hope people will learn that online harm can elicit as much pain as in-person harm: *"If you think that saying something bad to someone in person is bad, then you should also just assume the same for social media, it's not any different"(P26).*

Some participants expressed the belief that people need to educate themselves about the dangers of online harms, while others expressed a need for influencers and other celebrities

to get involved in public education about these issues. P8 experienced racist comments and thinks that people should educate themselves on the topic: *"I would like to see this community educating themselves more and uplifting Indigenous peoples rather than invalidating our experiences."* P7 hopes that celebrities or other public online figures can utilize their influence to *"empowering all types of individuals... and speaking up about the issues [online harm]."*

Education about online harm does not necessarily have to come from online sources. Some participants believe that it is particularly important for their school to educate students about online harms. For example, P18 was bullied online by her classmates and she wondered if the school could have prevented it with more education: *" If in eighth grade the school constantly talks to their students about social media bullying, online appropriation, using words correctly, not saying rude things online... if they do that from a younger age, then that would solve the issue from the beginning itself."*

4.4 Findings: Timelines of Needs

Our third research question asks about the order and timing people want to meet the needs. In the first several stages of the design activity, participants first identify the actions and relevant stakeholders for addressing harm, then reflect on their needs behind the choice. In the final stage, we asked participants to place the sticky notes representing different actions onto a timeline. Since the actions are attached to specific needs, we were then able to analyze the preferred temporal order of meeting the needs. We summarize the patterns that we found in Figure 4.2. Next, we present participants' perceived immediacy of the five identified needs: sensemaking, support and validation, safety, retribution, and transformation.

Need for sensemaking comes first

We find that when participants need to make sense of the situation, they usually do it before meeting other goals. As we discussed in 4.1, some participants were not sure if they overreacted, or they don't know how to proceed with addressing the harm. Thus, understanding offenders' motives and getting instructions for addressing harm are prerequisites for participants' next moves. The two outliers (P4, P16) chose to stop the continuation of harm before making sense of it.

Need for emotional support is dominant and happens at an early stage

The need for receiving emotional support and validation is dominant on the timelines. Many participants also wish to start to address it earlier in the process. About half participants place it as the first need to accomplish. P10 explained why she would like receive support first: *"I feel like the time after which the harm happens is when you're the most emotionally*

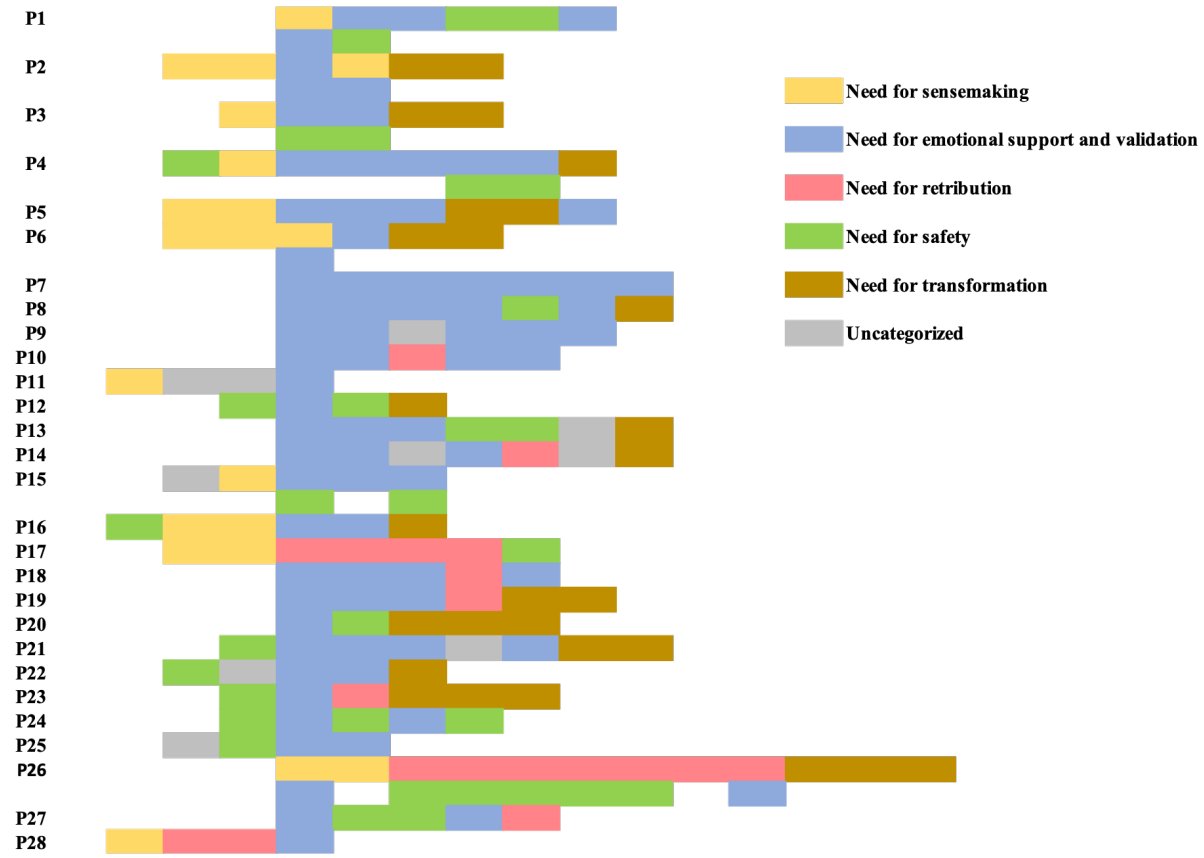


Figure 4.2: Participants' timelines to address the needs. Each rectangle represents an action on the timeline. We color-coded the actions according to which category the participant's need falls into – for an action that represents more than one need, we used stacked notes. We aligned the timelines according to the need for emotional support and validation.[1]

charged by this situation. So then, a comfortable space where you can ease in and grieve is pretty important." P27 believes that she has an urgent need to get "kind words and maybe a hug" from friends and family: *"When I saw it [the offending post], I was upset for the longest time. And I really considered ways of trying to avoid going to school and miss class so I wouldn't have to see them [offenders] again."*

Some participants hope to have a conversation with offenders to get their acknowledgement or stop the continuation of harm. We find that those participants usually need to receive emotional support before facing offenders. P3 explained that emotional support should come before confronting the offender: *"Because going straight to the offender and be like, 'Hey, what you did was wrong,' isn't going to happen if the survivor doesn't feel supported enough."*

Timing to meet safety needs depends on the types of actions

We noticed that some participants deal with harm that is ongoing, and hope to gain safety as soon as possible: *"once online harm happen, I guess the first thing to do is to stop it (P21)."* For participants who hope to stop the harm immediately, they usually rely on platform moderation (P4, P12, P16, P22, P24, P25). P4 thinks that removal of content as soon as possible can stop the spread of harm: *"Firstly, I want the Instagram moderators to delete the comments and give the Instagram users who posted the disrespectful comments some warning so these negative comments won't let more people to see it and let this discriminative mood to spread around individuals."* When participants hope to stop the harm through talking with offenders and gain their understanding (P1, P3, P8, P15, P27) it usually happens (or is finally achieved) at a later stage of the timeline. Some participants believed that it takes time for offenders to understand the harm, thus they hope to talk with offenders later: *"The offenders won't realize that what they've done was wrong at the time after they post the bad information. I will give them some space to let them understand what they've done was wrong."* Additionally, we talked about how P3 needs emotional support before talking with offenders (5.2).

As we mentioned in 4.4.3, sometimes participants are not sure how they can stop the harm. We find that some participants place actions to stop the harm at multiple stages of the timeline. For example, P24 hopes to first inform the platform and to get the offensive post removed, but he also hopes the school can tell offenders to stop their actions later. He admitted that ideally the offenders should stop the continuation of harm immediately, but *"It will take some time to talk to the students and all of that."*

Needs for retribution and transformation come last

We noticed that retribution is not the central goal for many participants. Less than half of our participants include the need for retribution in their timeline of addressing harm, and they usually place the need after meeting the need for sensemaking, emotional support and validation, and safety.

Finally, participants placed their need for transformation toward the end of their timelines. Participants told us that they hope the transformation of online environments is not only to address their harm case, but instead will prevent future harm: *"This one [the transformation need] definitely comes more at the end because this is more of trying to prevent future things from happening"(P3).* P2 believes that besides addressing past harm, people should "learn from this experience." Some participants indicated that the need for transformation are less for themselves and the harm they have experienced. Instead, they hope that fundamentally changing the online environment can benefit people they care about: "I hope that there are less and less victims that will be harmed." The need for transformation that our participants mentioned echoes calls for transformative justice [38, 88]. Rather than focusing on cases of harm individually, transformative justice seeks to reveal and address root

[1]For a grayscale version, please visit `https://applexiao.com/images/timelines.jpeg`

causes and cultures of violence and harm in society. Similarly, our participants discussed a longer term need to change the conditions that enabled harm to happen in the first place.

4.5 Discussion

Expansion of the scope of needs, stakeholders and actions in addressing online harm

Starting from victims' needs in the restorative justice literature, we identified five needs adolescents have for addressing online harm: sensemaking, support and validation, safety, retribution, and transformation. We also find that moderation actions, such as content removal or bans, meet some participants' needs for safety or retribution. This last finding aligns with existing research that shows that moderation grants participants safety [142] and can hold offenders accountable through retribution [145].

While content moderation is currently the major tool online platforms use to address harm, our findings suggest other ways platforms can help. In fact, participants mentioned online platforms' role in addressing all five needs we identified above. Participants believe that platforms can help them make sense of what has happened, validate their experiences of harm, and transform online environment to prevent future harm from happening. Further, participants proposed specific actions from online platforms that could help, such as providing instructions and advice on how to address harm, and showing their stance against harm. These proposed actions are useful steps towards concrete design solutions to online harm.

While platforms have a responsibility to their users to address harm, it is also clear from our interviews that victims need more than what platforms alone can offer. Social media companies, as well as current research, typically consider individual platforms or communities as the primary (or only) site for addressing harm. However, the restorative justice approach conceptualizes harm as an interconnected web of relationships, creating obligations for stakeholders (including offenders and members of related social circles) to address the harm collectively [184]. When we asked participants to choose which stakeholder(s) to engage with in order to address a harm, they often identify multiple stakeholders online and offline. This broader network of stakeholders requires us to think about online harm not as isolated incidents, but events that connect individuals experience with their relevant communities beyond a particular online platform. This recognition further emphasizes the importance of customization and flexibility when providing support [19, 145].

The multi-stakeholder perspective also reveals the importance of utilizing available social capital and resources that people have in their social circles [35, 112]. Kretzmann and McKnight explain that, "It is the capacities of local people and their associations that build powerful communities" [111]. For many victims we interviewed, their process of addressing harm involves stakeholders and resources from multiple social circles which may or may not directly relate to the community where a specific harm occurred. It is important to note

that the multi-stakeholder approach doesn't alleviate the responsibility of online platforms to protect their users. For example, one concrete suggestion is for platforms to directly point victims to the internal and external social resources they might need for addressing harm.

The involvement of multiple stakeholders can be particularly important for adolescents, our focal population. We find that their online harm experiences may happen between schoolmates, or even include an offline component at school or other extracurricular events. As a vulnerable group, adolescents often need the help of their parents or schools in dealing with harm. When online and offline harm intersect in schools, it creates a grey area of obligations between the school, the platform, and parents. Existing restorative justice practices for adolescents are often a collaborative effort between school and parents [72]. In the HCI and CSCW literature, researchers show how involvement of parents may benefit adolescents in dealing with online safety issues [177, 50]. Our findings support this line of research, and emphasize the importance of interpersonal and familial relationships for adolescents in order to prevent or respond to online harm.

Our research provides insight into how online platforms and communities might implement procedures that enact restorative justice values and processes. Participants identified many actions that can potentially embed restorative justice values. For example, support from bystanders, online community members, or society at large is important for acknowledging the harm and empowering victims. Instead of punishing the offenders to stop the harm, some participants explained that it might be possible to stop further harm through offenders' growth: offenders can learn from the harm and grow through conversations with victims and other online community members, or learn from schools and society at large. We believe these restorative measures are particularly important for the health and growth of adolescents who have experienced harm, or who caused harm to others. While restorative justice has demonstrated success in achieving those goals within schools [72], we believe online platforms are an important social context where even small, visible changes in our response to harm can have a positive effect on adolescents.

Understanding harm and need through a temporal perspective

When do online harm victims start to address the harm, and when does the process end? Our research shows that participants often deal with ongoing harm, or expect the harm to happen again even if it has temporarily stopped. While current research often conceptualizes harm as discrete incidents and designs solutions, it is important to take continuous or ongoing harm into consideration as well.

Participants' process of addressing harm often involves a series of actions over time. They also show preferences for addressing some needs in a particular order. Participants may hope to talk with offenders, but it's only after they gained emotional support; or they need to ensure safety before addressing other needs. Thus, it is important to consider both the relation of needs and the timing of meeting each need. For example, participants mentioned that sometimes they often need to make sense of what happened before deciding how to address the harm. Thus, besides designing solutions that focuses on the outcome, it is also

important to design ways to help users make informed decisions on how to address the harm. We also find that several participants see moderation as an efficient way to ensure safety right after a harm occurs. While much research examines *how* to moderate, our research shows the importance of *when* to moderate (i.e., moderation efficiency). Our participants expressed that they hoped for efficiency among both human moderators or automatic moderation tools. For this reason, we believe this is also relevant to the studies of labor and human moderators [131], and automation in moderation [61, 28].

In addition to short term needs and actions to address a specific harm, many participants explained that they wanted transformation of online environments in the long term. The broader need for transformation may be less apparent (and seem less urgent) compared to more immediate needs in an individual harm case. However, our interviews reveal that larger transformative changes are no less important to some individuals. In fact, many participants put it at the end of the timeline but describe the need as fundamental. In some ways, this is akin to wanting justice for a specific crime, but also wanting to change the laws and social norms that allow such crimes to regularly occur.

Since the focus of restorative justice is on interpersonal relationships [184], its effect on transformation may be limited. Participants' need for transformation include issues such as raising public awareness about online harms, and changing the platform's design to prevent future harms. These insights raise the potential of focusing on transformative justice in future research on online harm. Transformative justice aims to address structural conditions and root causes that enable harm to happen [38]. The challenge with taking a transformative justice approach is that unlike restorative justice, transformative justice is not well codified and does not have a set of established practices to build on. Instead, transformative justice is an open-ended, community based approach to understanding and addressing root causes while changing dominant, harmful cultures [88].

Reflection on our method: what we learn from victims' process of identifying needs

Our research process centers on the experiences of victims, and gives them agency to explain how they would want to deal with specific experiences of online harm. Of course, victims' proposed actions should not be conflated with specific, actionable implications that should be implemented directly. Instead, we argue that we should consider their suggestions as they relate to the values and norms within the relevant online communities and platforms.

There are many existing approaches to addressing online harms, and the most common actions such as banning, muting, and removing content are primarily punitive. Such punitive actions tend to focus on after-the-fact removal of offending content or the person who has broken the rules. In our research, some participants proposed such punitive actions, and some explicitly expressed the need to punish the people who hurt them. This should be expected when punitive approaches are the dominant way to address harm both on online platforms as well as in society more broadly [57]. The dominance of punitive actions creates

a dilemma for people who want to grant agency to survivors, but also aspire to the values of restorative justice. Thus, we must emphasize the difference between taking a survivor-centered approach, with one that is survivor-led. Taking a survivor-led approach means acting on exactly what a survivor of harm asks us to do; however, in a survivor-centered approach, we listen to survivors and work to meet their need to heal from the harm within the framework of established values. Therefore, it is important to establish agreement on shared values before implementing processes like restorative justice.

Our interviews used a multi-stage design task to capture participants' needs for addressing harm. We found that participants usually do not know how to address the harm immediately. They constantly reflected and came up with needs throughout the task, went back and forth in the procedures to add stakeholders or actions, and re-arranged the items on a timeline. In fact, many identified sensemaking as one of their needs in addressing harm — indicating that they may not know what has happened or know how to address the harm at the onset. We also found that the same action participants identified may relate to several different needs: for example, content moderation can satisfy retribution or safety; offenders' acknowledgement may either provide validation or act as a way to stop the continuation of harm (see Table 6.2). Therefore, identifying needs and actions is labor intensive and instead of expecting victims to provide us with direct, actionable solutions, it is important for us to provide time, support, and resources to them. Besides asking what actions they prefer, it is important to work to understand their underlying motives and needs. We believe that building on restorative justice values and processes can enable researchers, technologists, policy makers, and platforms to engage with victims of online harm in respectful ways that offer support and compensation as they work together toward designing new ways to effectively address online harm.

Limitation and Future Work

Our recruitment and interview processes have some limitations. First, the need-finding questions and examples, as well as the order they are presented in, have the potential to influence the responses of participants. Second, online harm is a broad topic; the cases participants shared do not cover all types of harm cases. Third, our participants may not be representative of all adolescents. Most participants are Asian. All participants are in their late adolescence (18-20) and study in a university in the United States. In future work, we plan to conduct large-scale surveys among adolescents to cover a wide range of online harm experiences and demographics. We also plan to use the survey to examine how people's experiences of harm relates to the needs, actions and stakeholders for addressing harm.

Our research has identified resources victims can mobilize in addressing harm, including stakeholders and what victims need from them. However, as we mentioned in 6.3, researchers and platform designers need to continue to work with victims to transform those ideas into design solutions. In particular, our research has shown potential of using restorative justice in addressing online harm. Applying restorative justice principles to online platforms is not easy or straightforward. Restorative justice processes can be time consuming, and may

require participation and effort from diverse parties including restorative justice facilitators, offenders and other community members[22]. In addition, a restorative justice process is often voluntary and consensual, thus it does not concern offenders who do not plan to engage [184]. Those issues present new challenges to the current online landscape which experiences insufficient moderation labor and expanding communities [62, 131]. While our research has showed the potential of online restorative justice, it is important to bear those constraints in mind in future design and implementation.

4.6 Conclusion

In this research, we identified adolescents' need for addressing online harm, including sense-making, support and validation, safety, retribution, and transformation. Our findings shed light on how online platforms may support victims beyond moderation, and show how we can design for victims' needs beyond the scope of online platforms, or short term solutions. Additionally, we see the potential of restorative justice in understanding and addressing adolescents' needs in online harm. How we may design for those needs and implement restorative justice principles in online platforms is challenging, yet important future work.

Chapter 5

RQ2: How do we meet online harm survivors' needs of sensemaking?

5.1 Introduction

Interpersonal harm, such as cyberbullying and sexual harassment, is a pressing issue on social media platforms [48, 170]. Online platforms primarily address these types of harm through content moderation, which focuses on punishing perpetrators through actions such as content removal or account banning. Survivors are often left out of decision-making in this perpetrator-centered framework. Research has found that survivors have unmet needs, including seeking advice, obtaining emotional support, and receiving acknowledgment and an apology from the person who caused the harm [181, 145, 121].

Given the growing scale and ramifications of online harm [48, 170], empowering survivors is a matter of societal and ethical urgency. In recent years, there has been growing interest within the fields of CSCW and HCI to adopt a survivor-centered approach by prioritizing the needs and agency of survivors and by designing tools and resources that empower them to take action on the harm they have experienced [157, 145, 121, 65, 44]. We build on this line of work and focus on empowering survivors in the process of *sensemaking*. Sensemaking is a crucial and central stage during which survivors gather information to understand the harm, recognize the resources available to them, and develop a plan of action to address their needs [174]. Research has found that it can be challenging for survivors to make sense of what they need and the actions to meet those needs within a perpetrator-centered content moderation process [182]. Survivors face challenges when seeking support in the sensemaking of harm, such as difficulty assessing the impact and severity of the harm [6, 65] and uncertainty about where to seek help [181, 165].

In this paper, we introduce *SnuggleSense*, a system designed to empower survivors through a structured sensemaking process. After survivors experience harm, SnuggleSense facilitates a process for them to understand the harm and develop a plan of action, especially in situations where immediate support is not available, or survivors are hesitant to reach out

due to the fear of secondary harm. To achieve this, we draw inspiration from a survivor-centered justice framework - restorative justice. Restorative justice is both a practice and philosophy of justice that prioritizes survivors' agency and needs in addressing harm. In recent years, CHI and CSCW researchers have applied restorative justice to comprehend online harm and provide support to survivors [145, 95, 74, 181], extending its application beyond traditional offline settings like the criminal justice system, schools, and workplaces [167, 180]. SnuggleSense follows this line of work and explores how we can apply restorative justice to expand the toolkit available to survivors for making sense of online harm.

SnuggleSense draws inspiration from two restorative justice practices: pre-conference and circles. First, SnuggleSense guides survivors through reflective questions inspired by the pre-conference process, where survivors work with a trained facilitator to process harm, identify needs, and develop an action plan [184, 181]. Second, SnuggleSense incorporates the social support aspect of circles by offering suggested actions from other survivors who have undergone similar experiences of harm. These elements are integrated into a design process facilitated by interactive digital sticky notes. The final outcome of the sensemaking process is a series of sticky notes arranged on a timeline, representing a step-by-step plan for addressing the harm in chronological order.

We compared how SnuggleSense facilitated survivors' sensemaking process to how they typically make sense of harm within the content moderation framework. We conducted a within-subject, controlled experiment where survivors developed an action plan for the harm they experienced using either SnuggleSense or by writing out the plan themselves. Our results indicate that participants found SnuggleSense significantly more effective in facilitating their sensemaking of harm. Participants appreciated the guidance provided by the structured sensemaking process and the agency enabled by the design. Additionally, we found that SnuggleSense encouraged survivors to adopt a community-based approach to addressing online harm.

We argue that SnuggleSense empowers survivors through a structured sensemaking process. It increases survivors' awareness of available resources and community, offering a pathway for addressing harm that emphasizes healing and restoration. We discuss the implications of SnuggleSense's design, including tailoring support to individual survivors, fostering a support community, ensuring safeguards for survivors as the system scales, and facilitating meaningful action following the sensemaking process.

5.2 System Design

The design of SnuggleSense is inspired by a restorative justice framework, which centers survivors' agency and needs and leverages community members as resources [184]. In this section, we first describe the guidelines we used to design SnuggleSense and how these guidelines have shaped the system's user flow. We then provide the implementation details of SnuggleSense. We conclude this section by reflecting on our positionality that informs the system design.

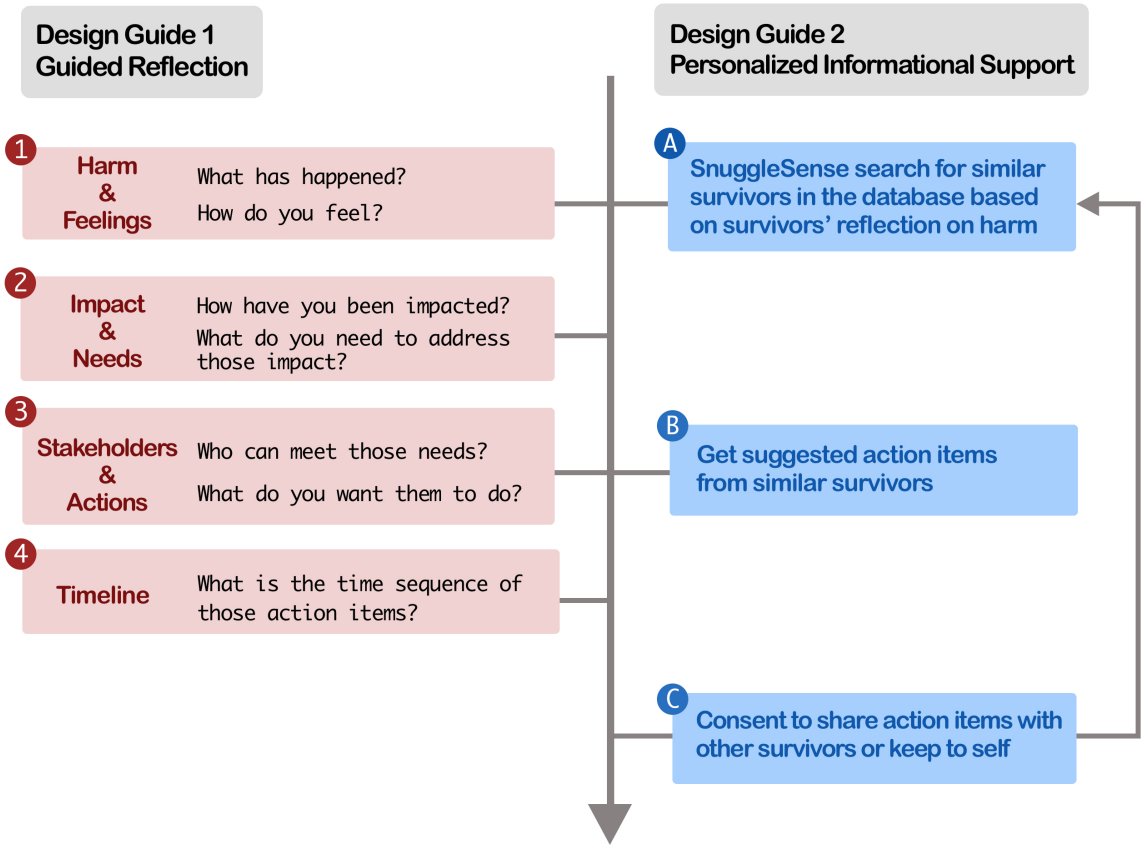


Figure 5.1: The Guided Reflection Process and Personalized Informational Support in SnuggleSense. SnuggleSense guides survivors' sensemaking process through a series of reflective questions inspired by restorative justice pre-conference. The questions prompt survivors to reflect on their experiences of harm, their feelings, the impact of the harm, their needs, and action plans to address those needs (steps 1-4). SnuggleSense also supports survivors' sensemaking process by providing them with personalized information. Based on each survivor's answer to the reflective questions, SnuggleSense searches for similar survivors in the database (step A) and recommends action items from similar survivors (step B). If consent is given (step C), survivors' action plans are incorporated into the database for making future suggestions.

Design Guide 3
Granting agency through a design process

1 Reflection on Harm & Feelings

Please take a moment to answer the questions below about your experience. Your responses will help us better understand the situation and provide suggestions tailored to your needs.

1. Please select all the options below that best describe the type(s) of harm you experienced:

- [x] Been called offensive names
- [x] Had someone try to purposefully embarrass you
- [ ] Been harassed for a sustained period
- [ ] Been sexually harassed
- [ ] Been stalked
- [ ] Been physically threatened
- [ ] Others

2 Reflection on Impact & Needs

Please take a moment to describe how the event has impacted you (less than 100 words).

It's disheartening to encounter such negativity and stereotypes in an online community meant for discussion and support.

3 Create Stakeholders and Actions

Now, create your own sticky notes below! You will have a chance to edit, delete, and receive suggestions later.

Stakeholder: Who can you reach out to for help or resolution? (one at a time)

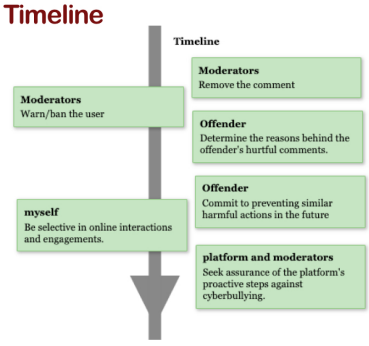
Action: What can they do? (one at a time)

Create an action item

My Action Plan

Moderators	Offender	Moderators
Remove the comment	Commit to preventing similar harmful actions in the future	Warn/ban the user

4 Timeline



A SnuggleSense search for similar survivors in the database based on survivors' reflection on harm

B Get suggested action items from similar survivors

My Action Plan

Moderators	Moderators	Offender	myself
Remove the comment	Warn/ban the user	Commit to preventing similar harmful actions in the future	Be selective in online interactions and engagements.
Edit Delete	Edit Delete	Edit Delete	Edit Delete

Discover action items from individuals with similar experiences

Here are some sticky notes from individuals who have encountered similar experiences. Feel free to **modify existing notes** or **incorporate their notes** to strengthen your action plan.

I'd like to explore data from people who have:

- [ ] Experienced the same type of harm
- [x] Faced harm in a similar location
- [ ] Encountered harm with a similar number of perpetrator(s)
- [x] Share a similar relationship with the perpetrator(s)

Update action items below

myself	A close friend whom I trust	Social Media Platform	the public
Be selective in online interactions and engagements.	listen and provide validation	Allow users to delete unauthorized posts	Advocate for support and understanding towards the victim.
Add to my action plan	Add to my action plan	Add to my action plan	Add to my action plan

< >

Back View more

C Consent to share action items with other survivors or keep to self

Willing to inspire others? Share your action plan below with fellow SnuggleSense users. If you prefer to keep your action plan to yourself, that's absolutely fine too. The choice is entirely yours, and we support your decision.

My Action Plan

Moderators	Moderators	myself
Remove the comment	Warn/ban the user	Be selective in online interactions and engagements.
Offender	**Offender**	**platform and moderators**
Determine the reasons behind the offender's hurtful comments.	Commit to preventing similar harmful actions in the future	Seek assurance of the platform's proactive steps against cyberbullying.

- ○ I will share my action plan.
- ○ I'd like to keep it to myself.

Figure 5.2: SnuggleSense Grants Agency Through a Design Process. The step number of the graph corresponds with Figure 5.1. SnuggleSense grants survivors agency through interactive sticky notes and a visual timeline for their action plans. Participants use these features to generate their action items (step 3), include suggested actions (step B), and visualize their plans on a timeline (step 4). These design activities serve to encourage survivors to exercise agency and creativity in exploring diverse ways to address harm and meet their unique needs. The screenshots in this figure illustrate the essential components of the system but do not encompass the entire interface.

Design Guides

Design Guide 1: Guided Reflection

Our first design guide is to provide survivors with guidance in their sensemaking of harm. When content moderation is the primary approach to addressing harm, survivors may not be encouraged to consider their role in addressing it and often struggle to envision solutions beyond this framework [182]. Survivors frequently need support when trying to make sense of the harm they've experienced [181, 107, 164].

SnuggleSense guides survivors through reflection questions inspired by restorative justice pre-conferencing, a step where a facilitator works with survivors to understand the harm and develop an action plan before engaging other stakeholders to reach consensus [184]. Salehi has drawn a comparison between the core questions the content moderation process and the restorative justice process ask [137]. In a content moderation process, these questions revolve around identifying reported content, determining its compliance with established rules, and deciding on appropriate actions such as removal, demotion, flagging, or ignoring. In contrast, the restorative justice process centers on different inquiries: Who has suffered harm? What are their needs? Whose responsibility is it to meet those needs?

SnuggleSense adheres to the questions above through the reflection process (Figure 5.1, step 1-4). In this process, survivors first engage in introspection, reflecting on the harm and the emotions it has created within them (Figure 5.1, step 1). Storytelling has a critical role in survivors' sensemaking, allowing them to recall and reconstruct their experiences from their unique perspective, thereby centering their emotions and experiences in the process [184]. Furthermore, storytelling can help resurface the details of the harm to aid in further reflections.

Once survivors have reflected on the harm and their emotional responses, SnuggleSense aids them in identifying the impacts of the harm and their associated needs (Figure 5.1, step 2). The restorative justice approach maintains that harm gives rise to impacts, and these impacts inform the needs of survivors [137]. Survivors often possess needs that conventional justice processes, such as punitive online content moderation, fail to address. These needs encompass elements like truth-telling, restoration, emotional support, and validation [184]. By encouraging reflection on impacts and needs before specifying concrete actions, this process enables survivors to conceive a wider range of potential strategies for addressing

harm that extend beyond conventional content moderation approaches.

After the reflection on impacts and needs, SnuggleSense guides survivors in the formulation of an action plan (Figure 5.1, steps 3-4). SnuggleSense's action plan comprises tasks assigned to various stakeholders (e.g., moderators, bystanders, family and friends). Restorative justice views the process of addressing harm as inherently multi-stakeholder, positioning survivors and perpetrators within their communities and recognizing the involvement of community members as vital contributors to the resolution process [184]. Research has also highlighted the complexity of addressing online harm, often necessitating the coordination of multiple stakeholders both online and offline [56, 181]. Following the online pre-conference procedure outlined by Xiao et al. [181], SnuggleSense breaks down the creation of an action plan into three parts: identification of stakeholders who bear responsibility or can offer assistance to survivors (Figure 5.1, step 3), identification of the actions these stakeholders can undertake to address the harm (Figure 5.1, step 3), and organization of these actions in a chronological sequence (Figure 5.1, step 4).

Design Guide 2: Informational Support from Survivors with Similar Experiences

SnuggleSense provides a unique online support system for survivors who may not have immediate human assistance after experiencing harm. It fosters a virtual community by storing and sharing action plans from survivors who consent to share their experiences with others. The design of informational support in SnuggleSense is inspired by restorative justice circles [184]. In these circles, survivors, along with other community members affected by the harm — including perpetrators, family, and friends — convene after the pre-conference to collaboratively formulate concrete action plans. Participation is voluntary, contingent upon resource availability and the willingness and commitment of the involved parties.

SnuggleSense aligns with this approach in a virtual setting, providing individual survivors with a platform to engage with and support one another. When immediate human support is not often available for survivors in their circumstances, how can we design an online system to connect survivors with the support they need? The system groups survivors with similar experiences together and shares their action items as suggestions. After a new user enters information about their specific harm case, SnuggleSense pairs them with survivors who have encountered similar experiences (Figure 5.1, step A). As users input their own action items, SnuggleSense presents them with action items from peers who have faced comparable situations (Figure 5.1, step B). Upon finishing their action plan, users have the option to choose whether to share their data with other survivors or keep it private (Figure 5.1, step C). This repository of action items becomes a tool for mutual help and understanding among users. To safeguard privacy, survivors are afforded the opportunity to review their data before deciding whether to share it or retain it confidentially.

Design Guide 3: Granting Agency Through a Design Process

SnuggleSense leverages spatial reasoning to support survivors, inviting them to present their action plan through interactive digital sticky notes and a visual timeline (Figure 5.2). Our approach to utilizing design to facilitate survivors' sensemaking draws inspiration from speculative design. According to Wakkary et al., design serves as a catalyst for exploring alternatives and redistributes the power of interpretation to the users [171]. They contend that design possesses the capacity to act as a bridge, connecting our present reality with an imagined, critically transformed perspective of our world. Moreover, Gerber underscores the notion that design artifacts can function as instruments for actualizing users' visions and igniting discussions and creativity around these concepts [60].

While addressing harm experienced by survivors differs from speculative design's focus on hypothetical scenarios, this design approach prompts survivors to contemplate ideals that transcend the constraints of the existing system. By drawing inspiration from speculative design, our aim is to use sticky note design activities to encourage survivors to exercise their agency and nurture a creative mindset, allowing them to explore a wide spectrum of approaches to addressing harm and meeting their unique needs.

Implementation of SnuggleSense

SnuggleSense is a web-based platform developed using a front-end stack that includes JavaScript, D3, jQuery, HTML, and CSS. On the back end, it is implemented in Python, leveraging the Flask framework, and is hosted on the Google Cloud Platform for data storage. The development of SnuggleSense followed an iterative design process, involving pilot testing and user feedback to refine and enhance its features. Next, we provide a detailed description of the SnuggleSense implementation.

Reflection: Harm, Feelings, Impact, Needs

In the initial phase of SnuggleSense, survivors are prompted to engage in self-reflection by documenting the harm they have experienced (Figure 5.2, steps 1). Initially, survivors provide a brief description of their experience in a text box. To facilitate this process, we have also included a set of multiple-choice questions aimed at encouraging survivors to consider various aspects of the harm they have endured. These multiple-choice questions not only stimulate survivors to examine their experiences from different perspectives but also serve as input for generating personalized recommendations at a later stage (Figure 5.2, step A). These questions cover four dimensions of the harm experiences that are relevant to their needs for addressing harm, including the nature of the harm, the location where it occurred, the number of individuals involved, and their relationship to the survivor. We selected these four dimensions of harm experiences based on pilot testing to determine which aspects participants found most relevant for identifying their needs. Following this, participants

further reflect on the impact of the harm and their needs for addressing it by writing in text boxes (Figure 5.2, step 2).

Create Action Items for Stakeholders

Following the reflective phase, SnuggleSense guides survivors in the creation of an action plan (Figure 5.2, step 3). We employ sticky notes to represent individual action items. Initially, we provide a sample action plan with example actions. Subsequently, we prompt survivors to compose their own action item, comprising a specific stakeholder and the corresponding action aimed at addressing the harm they have experienced.

Receive Recommendations from Survivors with Similar Experiences

After survivors have drafted their action plans, SnuggleSense offers support by presenting action item suggestions from other survivors who have encountered similar experiences (Figure 5.2, step B). Four suggestion sticky notes are initially presented to users, with the option to access more suggestions if desired. Users can integrate these suggestions into their existing action plans by clicking on the "Add to My Action Plan" button on the suggested sticky notes.

SnuggleSense offers relevant suggestions by grouping survivors with similar harm experiences. SnuggleSense calculates the similarity between the current user and existing users in the database based on multiple-choice questions they have answered about the context of harm (Figure 5.2, step A). In the database, a similarity score $S \in [0, 1]$ is stored for each pair of users. The similarity score is calculated as follows: For each multiple-choice question with n options, if both survivors either selected or did not select an option, $\frac{1}{n}$ is added to the similarity score. The total similarity score S_{ij} between two survivors i and j is the sum of the individual scores across all questions:

$$S_{ij} = \sum_{k=1}^{m} \frac{1}{n_k} \cdot I_{ik,jk}$$

In the equation, m is the number of questions, n_k is the number of options for question k, and $I_{ik,jk}$ is an indicator function that equals 1 if both survivors i and j selected (or did not select) the same option for question k, and 0 otherwise.

For each user, we identify three survivors with the highest similarity scores from the database and recommend their action items in a randomized order. Additionally, we provide users with four selection boxes representing different aspects of harm, allowing them to choose the most relevant suggestions based on their priorities. Once users finish drafting their action plans, we record the harm experiences and action plans they consent to share and store them in the database for future recommendations (Figure 5.2, step C).

Organizing Action Items Chronologically

Subsequently, SnuggleSense prompts survivors to organize their action items in a chronological sequence (Figure 5.2, step 4). Building on the research by Wong and Nguyen [179] and Xiao et al. [181], this step aims to help survivors visualize their action plans by listing the tasks in the order they intend to carry them out. Survivors have the flexibility to add additional action items as they construct their timelines.

Sharing Action Plans with the Community

In the final phase, SnuggleSense presents the completed action plan to survivors and inquires whether they would like to share it with fellow users of the system (Figure 5.2, step C). This feature encourages survivors to engage with a community of peers who can provide valuable insights and support in relation to their action plans with their consent. Participants can also choose to keep the action plan to themselves.

Positionality Statement

The authors of this paper have expertise in the fields of online harm and content moderation, with some having personal experiences with online harm. We reside in a society where punitive justice is the prevailing method for addressing harm, yet we acknowledge the merits of restorative justice, which emphasizes healing and restoration. The lead author has received training in restorative justice facilitation and has directly assisted online harm survivors using this framework. These experiences underscore our commitment to a survivor-centered approach in addressing online harm and exploring alternative approaches beyond the conventional punitive model.

While we draw inspiration from restorative justice with its survivor-centered nature and successful offline practice in helping survivors' sensemaking, our intent is not to advocate for it as the exclusive or prioritized way to address harm. Restorative justice's applicability is context-specific and varies depending on the nature of harm and individual circumstances [184]. We recognize the potential of different justice models. Traditional content moderation, as a punitive approach, has demonstrated effectiveness in stopping the continuation of harm and reducing re-offense [62, 79]. Addressing systemic issues such as sexism or discrimination requires a transformative approach aimed at rectifying the underlying structural problems [51]. Rather than advocating for a specific approach, our objective is to explore ways to empower survivors by drawing from alternative methods of addressing harm that are not traditionally adopted, thereby expanding the toolkit available to online harm survivors in addressing the harm they experience.

5.3 Evaluation

SnuggleSense aims to empower online harm survivors by providing a structured sensemaking process to enhance how they make sense of harm within the content moderation framework. To evaluate the system's effectiveness, we conducted a controlled experiment comparing survivors' action plans created with SnuggleSense to those developed through an unstructured sensemaking process, which reflects current approaches to making sense of harm

Controlled Experiment: A Comparison Between Writing Text and Using SnuggleSense

The experiment task was to produce an action plan for a harm case the participant experienced. We employed a within-subject design [41], allowing participants to compare two sensemaking approaches: an "Unstructured" condition and a "Structured" condition.

In the Unstructured condition, participants were asked to develop an action plan by directly writing a sequence of action items, each specifying a stakeholder and their corresponding actions to address the harm. This written approach served as the control condition, simulating the natural progression of an unguided sensemaking process. In the Structured condition, participants were asked to use SnuggleSense's guided sensemaking process to create an action plan, also structured by stakeholders and actions. Here, the action plan was presented on a visual timeline that utilized SnuggleSense's digital sticky notes to organize each item chronologically. Participants were allocated 15 minutes to complete the action plan in each condition.

We opted for a within-subject design rather than a between-subject design [41]. Our preliminary testing showed that participants often found it challenging to assess the effectiveness of their sensemaking or their sense of empowerment without being aware of alternative methods. A within-subject experiment allowed us to directly compare how participants' sensemaking experiences differed when applied to the same harm scenario [31]. To minimize potential priming effects, we presented the two conditions to participants in a randomized order [117].

We recruited individuals who had encountered harm in the past six months and asked them to reflect on an instance of harm that occurred within the designated timeframe set for this experiment. In our preliminary testing, we observed that individuals who had experienced harm a considerable time ago often had already engaged in substantial sensemaking and had developed a relatively stable perspective on how to address the harm. In some cases, they were no longer actively involved in the process of making sense of the harm. Recognizing that sensemaking is an ongoing and evolving endeavor that encompasses different phases [174], our recruitment criteria were designed to ensure that participants had experienced harm recently and were actively engaged in the sensemaking process.

Post-study Survey

After creating action plans in both conditions, participants were asked to complete a follow-up survey, where they assessed the sensemaking process from three key perspectives: the effectiveness of SnuggleSense in achieving its design objectives, the system's alignment with survivors' sensemaking goals, and the participants' ranking of SnuggleSense's individual features. A researcher was present to guide survivors through the evaluation process, prompting participants to provide rationales and posing follow-up questions after each section of the survey.

How the System Meets Its Design Goals in Sensemaking

The first part of the survey consisted of 5 rating questions. We asked participants to rate the two conditions on how well they performed along the following categories: guidance, support, agency, assistance in sensemaking, and empowerment. Participants gave a score of 1 to 7, 1 being strongly disagree and 7 being strongly agree. The first three categories, guidance, support, and agency, match the three key design guidelines of SnuggleSense: guided reflection, personalized informational support, and fostering agency through a design process. The fourth category assesses how effectively SnuggleSense achieves its primary objective of facilitating sensemaking. We included empowerment as the fifth evaluation criterion to explore how the sensemaking process might alter survivors' perceptions of their empowerment in the broader context of addressing harm.

How the System Meets Survivors' Goals in Sensemaking

Considering the varied experiences of harm that survivors have encountered, they may have distinct objectives in the sensemaking process [181]. Therefore, it is important for us to assess how participants achieve their individual goals within their specific contexts. In this step, we first asked participants to write down their goals in making sense of the harm. We then asked participants to rate how well the Unstructured and Structured processes met these goals respectively. Participants gave a score of 1 to 7, with 1 being the lowest and 7 the highest rating.

Ranking of SnuggleSense Features

Beyond assessing whether SnuggleSense achieves its intended goals, we are also interested in understanding how its various features contribute to meeting these objectives. At the end of the survey, we presented the list of features in SnuggleSense and asked participants to rank the top three that are useful to them. This approach helps to identify which elements of SnuggleSense are most instrumental in its overall effectiveness and enables participants to explain how specific features assist them.

Initial Data Collection

The initial collection of suggested action items in SnuggleSense was assembled from pilot testing sessions, during which participants documented action plans for addressing the harm they had encountered using SnuggleSense's individual reflection process (Figure 5.2, Step 1-4). The participants in these pilot tests were selected through convenience sampling [133] of people who have experienced online interpersonal harm in the past. We obtained consent from the participants to share these action plans for experimental use.

This initial dataset comprises contributions from 35 survivors, encompassing more than 200 action items. The action plans of pilot participants are based on a wide range of online harm experiences. 10 survivors had been called offensive names, 9 were intentionally embarrassed, 9 faced sustained harassment, 6 experienced sexual harassment, 1 was physically threatened, and 21 reported other types of harm. The incidents predominantly occurred on social media sites (31 participants), followed by texting/messaging apps (8), in person (2), personal email accounts (2), online gaming (1), forums/discussion sites (1), and online dating sites/apps (1), with 4 reports categorized as "other." In terms of the number of offenders, 14 survivors faced a single offender, 10 had 2-5 offenders, 6 had 6-10 offenders, and 5 had more than 10 offenders. The relationship with the offender varied: 17 participants were harmed by strangers, 8 by friends, and 12 by acquaintances. In Table 5.1, we presented the types of actions and stakeholders pilot participants mentioned in their action plans and their percentage in the initial dataset.

Safeguarding Participants in the Experiment

To ensure informed consent and user autonomy, users are informed of the sensemaking procedure through an introduction page before entering the system. After the sensemaking procedure, users are made aware of how their shared information will be utilized in the experiment and are given choices on whether to share their action items. Recognizing the severity of some online harm cases, we acknowledge that survivors may require additional support in addressing harm or during the sensemaking process. As a proactive measure, SnuggleSense includes a list of external resources, including non-profit organizations and a support helpline, prominently displayed at the top of the system. A researcher was present during the experiment to provide help when needed. Additionally, we mitigate the risk of exposure to inappropriate content through researchers' moderation. In both the initial dataset and any new data shared by participants while using SnuggleSense, researchers conducted a content review to ensure that it does not endorse violence or contain inappropriate material before granting access to others. These action plans are also anonymized, revealing only a general description of stakeholder types and their actions, rather than providing personally identifiable details. We used SnuggleSense in an experimental setting. As we scale the system with more survivors, we believe additional safety precautions will be necessary, which we discuss in section 5.5.

Table 5.1: The table shows the categories of stakeholders and actions mentioned by survivors in our initial dataset, collected prior to the experiment. The percentages represent the proportion of each category out of a total of over 200 action items.

Stakeholder Categories	Percentage in Initial Dataset	Action Categories	Percentage in Initial Dataset
Platform moderators	32.58%	Implement strategies to prevent future harm	14.77%
		Content moderation	9.09%
		Give advice	4.92%
		Help me understand the harm	3.79%
Offenders	24.24%	Understand the impact of their actions	7.58%
		Apologize	6.44%
		Explain their motivation	5.68%
		Change their behavior	3.41%
		Stop the continuation of harm	1.14%
Online community members	21.21%	Give emotional support	8.71%
		Raise awareness	6.82%
		Report inappropriate comments	3.41%
		Give advice	2.27%
Family and friends	17.05%	Give emotional support	10.98%
		Give advice	6.06%
Myself	4.92%	Be more cautious in the future	2.27%
		Communicate with offenders	0.76%
		Ignore, block, delete, leave	0.76%
		Report	0.38%
		Self-care	0.38%
		Communicate with people I trust	0.38%

Participant Recruitment and Experiment Setup

We recruited 32 participants to conduct the within-subject study from July to August 2023. We recruited these participants using a campus-wide recruiting system at a West Coast university in the United States. We randomly assigned participants to do the Unstructured condition or Structured condition first.

Participants have an average age of 20.61, with 22 Female, 7 Male, and 2 Non-binary. There are 16 Asian, 7 White, 1 Latino, 1 African American, and 6 Mixed. One participant chose not to reveal their demographic information. Regarding their experiences with online harm, 24 participants had been called offensive names, 18 were intentionally embarrassed, 8 faced sustained harassment, 5 experienced sexual harassment, 4 were stalked, 4 were physically threatened, and 6 reported other types of harm. Most instances occurred on social media sites (20 participants), followed by forums (7), messaging apps (6), online gaming (6), and online dating apps (3). Additionally, 6 incidents had an in-person component. The number of perpetrators varied: 12 participants had a single offender, 16 had 2-5 offenders, 2 had 6-10 offenders, and 2 had more than 10 offenders. The majority (24 participants) were harmed by strangers, 6 by acquaintances, and 4 by friends.

Participants completed the task remotely with their personal computers. A researcher was present via zoom to provide an introduction of the study in the beginning and guide the participant in the follow-up survey, and participants independently completed the task to create action plans in between. This entire process spanned approximately 50 minutes. Participants received a compensation of $25 US dollars. The study was approved by our University's Institutional Review Board.

5.4 Result

In this section, we are looking at how well both the Unstructured and Structured conditions performed in four important areas: how the system meets its design goals, how the system meets survivors' goals in sensemaking, identifying the most useful features of SnuggleSense, and comparing the action plans in both conditions.

Design Goals: Guidance, Support, Agency, Sensemaking, and Empowerment

We conducted a two-tailed, paired t-test to assess the differences in ratings between the Unstructured and Structured conditions, focusing on the system's design goals of guidance, support, agency, sensemaking, and empowerment. We presented the data in Figure 5.3. The findings indicate that the Structured condition received significantly higher ratings in guidance (Unstructured: M = 4.94, SD = 1.24; Structured: M = 6.13, SD = 0.91, p ¡ .001), support (Unstructured: M = 4.56, SD = 1.46; Structured: M = 5.75, SD = 1.05, p ¡ .001), sensemaking (Unstructured: M = 4.75, SD = 1.46; Structured: M = 6.06, SD = 0.98, p ¡

Figure 5.3: The average ratings participants gave to the 5 design goals in the Unstructured and Structured conditions. The rating scale is from 1-7, where 1 indicates strongly disagree and 7 indicates strongly agree. Two-tailed t-tests; standard deviations in parentheses, * p ¡ .05, ** p ¡ .01, *** p ¡ .001. Bars signify standard error.

.001) and empowerment (Unstructured: M = 4.91, SD = 1.55; Structured: M = 5.94, SD = 0.91, p ¡ .01). There was no statistically significant distinction between the two conditions in terms of agency (Unstructured: M = 5.50, SD = 1.30; Structured: M = 5.53, SD = 1.11, p = n.s.). Next, we explored the rationale participants provided for their ratings across the five criteria.

Guidance

Participants favored the Structured condition for its effectiveness in providing guidance. They expressed appreciation for the step-by-step procedures: *"It's very organized, very efficient in terms of guiding me to resolve the incident" (P15).* Moreover, participants highlighted that SnuggleSense was instrumental in breaking down complex emotional and cognitive states into manageable components, enabling them to address one problem at a time: *"I think my thoughts and emotions are such a busy place. . . I do feel like it [SnuggleSense] just breaks it down a little more and helps me address one problem at a time and from the start to finish with the prompts" (P32)."* One participant described the sensemaking process with SnuggleSense as "hand-holding": *"I could write down my problems as a journal and follow through very effectively...it kind of hand-holds you through the process" (P6).*

Support

Participants reported experiencing greater support while using SnuggleSense. The source of support most participants mentioned is the recommended action item from others who had gone through similar situations, which alleviated their feelings of isolation: *"It [SnuggleSense] definitely made me feel like I wasn't alone, whereas [the Unstructured condition], it felt very like on my own and like not connected with anyone" (P14).* In addition, the features to facilitate sensemaking, such as the creation of sticky notes and timelines, were also cited as offering additional support, in contrast to the Unstructured condition where participants felt they were left to navigate independently: *"It was a very supportive system, like the input from other people who went through the same thing and then also being able to make the timeline and sticky notes really easily... but [the Unstructured condition] didn't have any support for that. It's just, it was all on my own" (P20).*

Agency

Our research did not identify significant differences in the ratings of agency between the Unstructured and Structured conditions. When examining their rationales, we discovered that participants experienced agency through different pathways. In the Structured condition, agency emerged from the diverse approaches available for exploring harm and the sense of control facilitated by the design features: *"There was definitely a lot more freedom with the sticky note process. With the interface of like when we had to put it on the timeline, I liked that you could actually just drag them anywhere... It felt more in my control, so I like that a lot" (P14).* Conversely, in the Unstructured condition, agency resulted from participants owning the process themselves, free from the need to adhere to prescribed steps: *"You have more freedom or like flexibility in terms of how you want to approach it [the sensemaking process]...there is more flexibility in the sense that gives you more options to practice the actual process to address the harm" (P31).* Some survivors preferred to navigate the sensemaking process independently rather than seeking suggestions from others: *"I'd think through my own actions, more myself versus the sticky note condition kind of was looking into what others have done" (P29).* Participants also emphasized that it was not an either-or choice, as both conditions allowed them to retain agency over the action plan: *"I thought they both provided agency because we owned the action plan in both cases" (P29).*

Sensemaking

Participants found the Structured condition more effective for sensemaking. Participants constantly cite the design features and the structure of SnuggleSense as helping with sensemaking, such as the timeline and the guided reflection process. The Unstructured condition, in contrast, was seen as more open-ended and less directive. To many participants, SnuggleSense provides a roadmap for thinking about the harm, not only at its occurrence but also in planning for the future:

> *"The [Unstructured] one, it didn't really like, tell me how to address it. I just kind of wrote what happened to me and that was it. But with the sticky note, it was really helpful, like the whole timeline thing to actually like reorder my steps and like, see what I would do in that process... really helped just to like, delineate how I'm going to address this in the future" (P26).*

Empowerment

The Structured condition was seen as more empowering than the Unstructured condition, especially because it allowed participants to see the thought processes of others who had faced similar issues: *"A lot of the time people are hesitant to involve with authority figures because they feel like their problems aren't worth it. So to see that other people were having the same thoughts of me was empowering" (P18).* Some participants also expressed that agency given by the design features is empowering: *"The sticky note one felt empowering because I felt that I could delete stuff or add stuff and then seeing what other people wrote and then the timeline, like being able to think through what I would want to do first and then to move forward with and having a timeline of things to do was just empowering" (P16).* The same participant noted that the Unstructured condition was also considered empowering, but in a way that left participants to create their own unique solutions: *"[The Unstructured condition] was empowering in a different way where I created my own solutions completely and then just had some sort of framework to do something but it was in a different way, I guess" (P16).*

Self-defined Goals

In the follow-up interview, participants were asked to establish self-defined goals for their sensemaking of harm and rate the effectiveness of the two conditions on each goal using a scale of 1 to 7, with 1 being the lowest and 7 the highest rating. Table 5.2 presents a summary of the key metrics along with the example goals articulated by participants. We conducted qualitative coding of participants' self-defined goals and the majority of them could be categorized into six categories: (1) Understand and assess the harm itself, (2) Come up with an action plan, (3) Manage emotions or engage in self-care, (4) Specify actions by stakeholders (including actions by themselves), (5) Prevent harm from happening in the future, (6) Actively address the harm.

To assess how well the two conditions aligned with participants' self-defined goals, we conducted a two-tailed, paired t-test on the arithmetic mean ratings of the self-defined goals each participant gave to the two conditions. The findings indicate that the Structured condition received significantly higher ratings in meeting participants' self-defined goals (Unstructured: M = 4.35, SD = 1.56; Structured: M = 5.56, SD = 1.29; p ¡ .001) [1].

[1]We used the arithmetic mean ratings for the t-test because of the diverse range of goals articulated by each participant. It is important to note that participants were not explicitly instructed to weigh their goals, and, therefore, the assumption of equal weighting is a limitation of this analysis.

Table 5.2: In the follow-up survey, each participant has written down a number of self-defined goals for their sensemaking of harm. The graph presents the major categories of goals participants mentioned, the percentage of participants who mentioned each category, and examples of the goals that participants created.

Categories	Percentage of participants mentioning the category	Examples from participants
Understand and assess the harm itself	40.63%	Understanding why the person wanted to cause me harm, identifying who's at fault
Come up with an action plan	31.25%	Developing structured actions for the incident, thinking about how to move forward
Manage emotions or engage in self-care	43.75%	Understanding it is not my fault, separate myself from the situation
Specify actions by stakeholders (including actions by themselves)	43.75%	Connecting with a support network in order to receive help, discussing with loved/trusted ones like family
Prevent harm from happening in the future	28.13%	Learn to respect my own boundaries, move forward and prevent similar situations from happening
Actively address the harm	6.25%	Addressing the harm that was taken, working through possible ramifications/consequences of the harm

The Most Useful Features

Our analysis identified the three features of SnuggleSense that participants found most useful: receiving recommendations (mentioned by 65.63% of participants as top three), sorting action items on a timeline (mentioned by 59.38% of participants as top three) and creating stakeholders and actions (mentioned by 53.13% of participants as top three). Receiving recommendations emerged as the most frequently cited useful feature, chosen as the single most useful by 37.50% of participants.

Next, we delved into participants' rationales for selecting the top three most useful features. In addition, we discussed how sharing action plans with others in SnuggleSense enhances a sense of connection and collaboration among its users.

Receiving Recommendations

Participants expressed that the recommended actions provided by SnuggleSense were pertinent to the challenges they were facing. As one participant explained, *"Suggestions they give me, they're so tailor-made to kind of similar problems I'm dealing with and it helps inspire me to different ways to address the situation" (P32).* Participants appreciate the offered insights from individuals with similar experiences, which inspired them when devising their own courses of action: *"I think a huge point in this thing's favor is that it shows you how other people dealt with the same issue and that gives you a lot more ideas than just trying to think on your own" (P6).*

Moreover, participants highlighted the emotional validation they derived from reading about others' action items. This validation reinforced the notion that they were not alone in their struggles, as expressed by one participant: *"To know that other people are also feeling the same way or similar ways as you are is very validating" (P3).*

Sorting Action Items on a Timeline

Participants appreciated the ability to sort their action items on a timeline, as it helped them organize their thoughts and visualize their action plans. The timeline served as a practical tool for planning actions, whether they were to be taken in response to an imminent situation or during a potential recurrence of the harm.

> *"My top priority when something like this happens is usually to assess it in my brain, sort of rationally, and decide where to go from there. What's my next course of action? So sorting recommendations down on a timeline really helped me order my thoughts." (P17)*

> *"It was very useful to visualize my plan of actions that I would take, say like if this were to happen again ... I'd immediately be able to take action instead of just kind of like be in shock." (P26)*

Creating Stakeholders and Actions

By identifying the various stakeholders involved, participants were better equipped to understand the complexity of the situation. Participants indicated that the step helps them to identify root cause of harm and alleviate their self-blame: *"Not doing that [creating stakeholders and actions] makes all the pain jumble up into one, and it can cause very ineffective or not healthy ways of coping with the problem if you're not identifying what really is causing you pain" (P7).*

Moreover, creating stakeholders and actions allowed participants to assign responsibility for addressing the harm. Rather than merely feeling distressed about the situation, they could proactively identify individuals who could instigate change: *"Instead of just like, you know, feeling bad about the situation, you can actually be like, okay, this person can actually change something, and then like thinking about what could be done is pretty helpful" (P4).*

Sharing Action Plan with Others

It is worth noting that while sharing their action plan with others is not an integral part of survivors' sensemaking process of their own experiences of harm, we found that it introduced a sense of community and fostered an empowering give-and-take dynamic. Participants emphasized that sharing their action plans made them feel like they were part of a community. In contrast to the solitary act of simply writing down their thoughts, sharing created a sense of connection and contribution: *"It makes you feel part of a community. And when you're just writing things down, you don't really get that, but when you share it, you kind of feel like you're contributing back and helping more people in the future" (P6).* This mutual exchange of knowledge and support was perceived as therapeutic: *"I liked that not only was I able to use other people's recommendations, I could also submit my own for someone else that might need it in the future. So it's not just a take, it's a give and take... That's therapeutic" (P7).*

Action Plan

In this section, we examine how participants formulated their action plans in both the Unstructured and Structured conditions. We compared the types and proportions of stakeholders and actions included in participants' plans across the two conditions. Additionally, to assess how participants used SnuggleSense's recommendations, we analyzed the types and proportions of recommended action items that participants incorporated into their plans.

Number of Stakeholders and Actions

We counted the number of distinct stakeholders and action items participants mentioned in the action plan of both conditions. The data revealed that participants incorporated significantly more stakeholders in their action plan when using the Structured condition (M = 4.34, SD = 1.64) compared to the Unstructured condition (M = 3.16, SD = 1.08), p ¡ .001. In addition, participants formulated significantly more action items in the Structured condition (M = 6.25, SD = 2.87) than the Unstructured condition (M = 4.50, SD = 2.23), p ¡ .001.

Categories of Stakeholders and Actions

Table 5.3 and Table 5.4 present the main stakeholder and action categories mentioned by participants in both the Unstructured and Structured conditions. These tables also provide the percentage of participants who referenced each category in each condition and the percentage point difference between the two conditions.

In the Structured condition, participants displayed a greater tendency to involve various stakeholder types (e.g., family and friends, platform moderators, and online community members) rather than assigning actions to themselves (as shown in Table 5.3). When analyzing the shift in action types from the Unstructured to the Structured condition, we see a reduction in instances of self-directed problem solving, such as opting for actions like "Ignore,

Table 5.3: Comparison of Stakeholder Types Mentioned by Participants in Unstructured and Structured Conditions (ordered by percentage point difference). The first column presents the primary stakeholder categories identified by participants in both conditions. The second and third columns depict the percentage of participants who mentioned the stakeholder category in each condition. The final column illustrates the percentage point difference between the two conditions.

Stakeholder Categories	Unstructured	Structured	Percentage point difference (Structured minus Unstructured)
Myself	68.75%	53.13%	-15.62%
Offenders	75%	81.25%	6.25%
Platform moderators	78.13%	90.63%	12.50%
Online community members	25%	53.13%	28.13%
Family and friends	37.50%	68.75%	31.25%

block, delete, leave" (refer to Table 5.4). Simultaneously, there is a significant increase in participants seeking explanations from offenders (a 28.13% increase) and soliciting advice (a 34.38% increase) or emotional support (a 28.13% increase) from online community members.

The data suggests that the Structured condition leads participants to consider a more diverse and inclusive range of stakeholders, shifting focus from a self-centric approach in the Unstructured condition to a more community- and network-centric perspective. It also promotes the consideration of approaches beyond content moderation as the sole means of addressing harm, encouraging actions that delve into the underlying causes of harm and seek support from various sources. This shift is echoed in the reflections of participants:

> *"There's something in [the Structured condition] that you see, it [the harm] is a systemic problem. It's not just some random one bad guy in the world that wants to send harmful messages. There could be a lot of other people involved that can, you know, make this issue better." (P2)*

> *"When I was just writing things down [in the Unstructured condition], I wasn't thinking as much about who was at fault, but I think [the Structured condition] helped to clarify that a bit more and just understand that it wasn't really on me for what happened." (P13)*

Adopted Suggestions

In the Structured condition, a majority of participants (81.25%) adopted SnuggleSense's recommended action items into their own action plans, and 42.22% of participants' action

Table 5.4: Comparison of Action Categories Mentioned by Participants in Unstructured and Structured Conditions (ordered by percentage point difference). The first and second columns present the primary stakeholder and action categories identified by participants in both conditions. The third and fourth columns depict the percentage of participants who mentioned the action category in each condition. The final column illustrates the percentage point difference between the two conditions.

Stakeholder Categories	Action Categories	Unstructured	Structured	Percentage point difference (Structured minus Unstructured)
Myself	Ignore, block, delete, leave	53.13%	28.13%	-25.00%
Family and friends	Give emotional support	28.13%	40.63%	12.50%
Offenders	Stop the continuation of harm	31.25%	43.75%	12.50%
Platform moderators	Content moderation	72%	87.50%	15.63%
Family and friends	Give advice	19%	40.63%	21.88%
Offenders	Explain their motivation	6.25%	34.38%	28.13%
Online community members	Give emotional support	9.38%	37.50%	28.13%
Online community members	Give advice	6.25%	40.63%	34.38%

items were derived from SnuggleSense's recommendations. This aligns with participants' identification of recommended action items as the most valuable feature.

Furthermore, 65.63% of participants added new categories of stakeholders they had not previously considered. Specifically, 40.63% added online community members as a new category, 25% added themselves, 15.63% added friends and family, 15.63% added platform moderators, and 12.50% added the offender.

Additionally, 81.25% of participants added new actions to existing stakeholders or new stakeholders. The top three new categories of actions participants added were for platform moderators to implement strategies to prevent future harm (40.63%), for online community members to give emotional support (28.13%), and for offenders to explain their motivations for conducting harm (25%).

5.5 Discussion

In this paper, we introduced SnuggleSense, a system designed to empower survivors through a structured sensemaking process. Our evaluation demonstrates its effectiveness in enhancing survivors' sensemaking. In this section, we reflect on these findings and explore their implications for addressing online harm and offering support to survivors in online spaces with social computing systems. We first argue that the sensemaking process enabled by SnuggleSense has the potential to empower online harm survivors, granting them agency and power in meeting their needs to address harm. Next, we explore how SnuggleSense opens up a restorative justice pathway for harm resolution. In the end, we reflect on our design lessons, future work, and limitations.

Sensemaking as a Process towards Survivor Empowerment

Empowerment is the process by which individuals and collectives gain control over issues that affect them [129, 52]. In the context of online harm, the empowerment of survivors can be seen as the process where survivors gain control over how to address the harm they experience. Recent HCI and CSCW work has explored tools to empower survivors, ranging from those facilitating social support [108, 19] to those aiding in the collection and documentation of evidence [158, 65]. Our research adds to the line of work by empowering online harm survivors through a sensemaking process. In the following, we discuss how the sensemaking process in SnuggleSense empowers survivors from two perspectives: first, by providing a structure for sensemaking, and second, by making survivors aware of their support communities and the resources available to them.

Empowerment through a structured sensemaking process

The structured sensemaking of SnuggleSense enables survivors to establish a clearer and more actionable connection between their goals in addressing harm and the means to achieve them. Our research found that SnuggleSense provides significantly more guidance than writing out an action plan, and sorting actions on a timeline is one of the most important features perceived by participants. Zimmerman argues that empowerment occurs when individuals can perceive a direct correspondence between their goals and how to attain them [186]. The structured sensemaking process of SnuggleSense, with its guided reflective questions and the visualization of action plans on the timeline, helps survivors gain a deeper understanding of their experiences and how to address their needs.

The structured sensemaking process in SnuggleSense enhances the knowledge survivors need to address the harm they experience. Our results show that participants using SnuggleSense developed action plans with a wider range of actions and stakeholders. A key aspect of psychological empowerment is increasing awareness of available actions and strengthening problem-solving skills [143]. Through its guided reflection questions and recommended

actions, SnuggleSense provides a framework for survivors to reflect on their experiences and needs, expanding their understanding of potential actions to address the harm they face.

Empowerment through awareness of communities and resources

Sense of empowerment can be enhanced by sense of community [32], as well as the ability to identify those with power, resources, and connections to the issue of [186]. SnuggleSense can empower survivors by fostering awareness of the communities they are part of and the support and resources available to them. The community SnuggleSense introduces are two-fold. First, it encourages survivors to consider their social circles, such as family and friends, or the online communities where the harm occurred. Participants highly valued the function of identifying stakeholders and their actions in SnuggleSense, mentioning significantly more stakeholders and their actions compared to the Unstructured condition in their plans.

Second, SnuggleSense facilitates survivors to find inspiration and validation in other survivors' experiences. Survivors rated receiving recommendations as the top useful feature. Further, survivors are also empowered by sharing and contributing to other survivors who use SnuggleSense. Participants highly valued the ability to share action plans and inspire others. It gives them a sense of community and they gain agency and control through giving back to the community. Zimmerman believes that being involved in community organizations can exercise a sense of competence and control [186]. Survivors derive strength from one another, and the willingness to share their plans with others demonstrates the platform's potential to foster a sense of community among survivors.

An empowered individual is essential for empowered communities [103, 186]. In addition, connecting with more stakeholders facilitates community empowerment by raising awareness of a problem's existence and negotiating common goal [106]. Besides aiding survivors in addressing current harm, we envision SnuggleSense as a tool that also serves to educate and empower the community in the long run. SnuggleSense offers a sensemaking framework that can be applied to future instances of harm experienced by a survivor or others.

A Restorative Justice Approach to Addressing Online Harm

Our results also indicate how a restorative justice pathway empowers survivors to consider community-based harm resolutions and prioritize restoration and healing. Our research indicates a shift in survivors' responses when utilizing SnuggleSense, involving a broader array of online and offline stakeholders, including family, friends, and online community members, in the process of addressing harm. In addition, survivors move away from individual efforts such as blocking, muting, or solely relying on punitive measures of moderators to understand the motivations behind harm or seek emotional support.

These observed shifts align closely with the recommendations put forth by the researcher community, emphasizing the need for designing interventions that prioritize survivors' healing and restoration needs [145, 121, 181, 65, 157]. Moreover, our findings resonate with the work of researchers who embrace a community-based approach to addressing harm. For

instance, Squadbox employs "friend-sourcing" to empower survivors [108], while Heartmob relies on online community members to provide assistance [19]. SnuggleSense builds upon this foundation by providing structured guidance and asynchronous support, allowing users to draw inspiration from others' journeys and find validation without the need for continuous online support.

Importantly, our research underscores the potential of restorative justice principles in achieving these transformative shifts. Restorative justice encourages people to identify the root cause of harm and emphasizes support and healing instead of punishing the perpetrators [184]. It locates harm in communities and argues that community members have a stake in addressing the harm [184]. It is worth noting that SnuggleSense did not explicitly dictate the stakeholders or actions involved; rather, these results emerge organically through the empowerment of survivors and their agency in the sensemaking process. SnuggleSense joins other work and shows how restorative justice provides a potential pathway in online harm-resolution that complements the current approach [145, 95, 181].

Design Insights and Future Work

SnuggleSense demonstrates how a social computing system can support online harm survivors in the sensemaking process. Our experiments with 32 participants highlight SnuggleSense's potential to scale and assist a broader range of survivors. Moreover, we believe SnuggleSense's design provides valuable insights for developing future social computing systems that support survivors, particularly by facilitating sensemaking and promoting community awareness. In this section, we reflect on the design lessons learned from deploying SnuggleSense in an experimental setting, with the goal of informing the design of future social computing systems.

Tailored Support to Survivors

In SnuggleSense, we provide survivors with informational support by suggesting relevant stakeholders and actions. This is achieved through algorithms that assess the similarity of survivors' responses to multiple-choice questions about their harm experiences. Future systems have the potential to further refine these recommendation mechanisms to better tailor support to survivors.

The similarity between survivors can be measured using diverse metrics. Recent research found that online harm survivors' needs can be influenced by various factors, including personal traits (e.g., demographics [145, 146], role in society and culture [172]), past experiences with harm [145], or the context of harm (e.g., their relationship with the perpetrators [172], the time span of harm [164]). These factors present opportunities for tailoring suggestions to survivors. In addition, these aspects may influence survivors' needs differently and hold varying degrees of importance for different individuals. In future work, we plan to conduct large-scale surveys to explore how participants harm experiences and their personal traits influence their needs differently.

When providing personalized recommendations, it's important to balance guidance with agency. We acknowledge that while providing guidance can empower survivors, it can also limit their agency. Our research revealed no significant difference in how the Structured and Unstructured conditions provide a sense of agency to survivors. When participants explained their preferences, some found the Structured condition offered more freedom and control by allowing them to take ownership of the design process. In contrast, others appreciated the Unstructured condition as it required them to think more deeply about their actions without external guidance. In SnuggleSense, we chose to let survivors initially reflect on the harm independently before providing suggestions. Finding the optimal balance between these two objectives is an important challenge to explore in future work.

Nurturing a Support Community among Survivors

SnuggleSense highlights the potential to foster mutual aid communities among survivors of online harm. In traditional online support groups, help often comes from bystanders or community members who may not share the survivors' experiences. Prior research has explored how individuals seek support on social media platforms, such as using Reddit throwaway accounts [5] or engaging with the #Depression tag on Instagram [7]. However, these approaches encounter challenges. Survivors may experience secondary harm from individuals who lack a deep understanding of their experiences [165]. Furthermore, differing perspectives on how to address harm—often from bystanders or external stakeholders—may not align with survivors' actual needs or desires [182].

SnuggleSense offers an alternative by highlighting the essential role survivors can play in addressing harm within their own community. Through mutual exchanges, survivors share contextually relevant advice, affirmation, and validation. This aligns with Fraser's work on the value of self-paced, internal discussions among marginalized groups [55]. By creating a space for survivors to share their experiences and action plans, SnuggleSense empowers individuals to explore and affirm unique strategies for addressing harm—strategies that are often overlooked by traditional content moderation systems.

SnuggleSense invites us to explore the potential of creating systems that foster survivor-led support communities. SnuggleSense facilitates the asynchronous exchange of action plans, enabling survivors to find informational support even when external resources are unavailable. Our findings show that participants value this reciprocal dynamic: receiving suggested actions was the most appreciated feature, and survivors felt a sense of reward for contributing to the community. Thus, future systems can consider supporting more varied forms of interaction within survivor communities. Survivors could validate others' proposed actions, share insights, or even return to the platform to provide updates on their progress after addressing harm. By fostering a cycle of support, such systems have the potential to nurture a supportive network that leads to community empowerment.

Safeguarding Survivors

Participants in our study used SnuggleSense in an experimental setting. Deploying systems that support survivors at scale requires consideration of additional safety measures. Similar to prior online support communities, it will require content moderation to identify and remove inappropriate content [19, 5, 7]. Many survivors may have prior experiences of harm, making it crucial to adopt a trauma-informed design approach [33, 147] to safeguard them from secondary harm.

Survivors' experiences and needs are individualized and may change over time [181, 174]. When using algorithms to provide personalized suggestions, it is important to assess how these algorithms influence survivors' decision-making processes and whether they deliver recommendations that cater to survivors' needs and the system's goals [183, 90, 141]. In addition, the system should continually review and update the platform's security and safety measures, providing survivors with the ability to modify or revoke their consent as their needs evolve [77].

Beyond Sensemaking: Taking Actions

While SnuggleSense focuses on the sensemaking stage, taking action is a crucial component in achieving empowerment in practice [186, 121, 19]. There are challenges for survivors to implement the actions they propose. SnuggleSense provides avenues for addressing harm that are not traditionally applied, making it hard for survivors to envision the effectiveness of those alternatives [182]. Therefore, the motivation to act on the action plans can be directly linked to the availability and accessibility of resources for survivors to act on their newfound understanding. We believe that it is essential to pair the improvement of survivors' sensemaking process through SunggleSense with the allocation of resources and the creation of supportive conditions for survivors to act. This includes creating ways to assemble relevant stakeholders and resources to assist survivors [158, 108, 19, 65], or changing societal attitudes toward addressing harm [182]. In future research, we aim to conduct longitudinal studies to explore how survivors continue to engage with their action plans developed through SnuggleSense, with a focus on identifying, designing, and consolidating resources to support survivors in the ongoing process of addressing harm. We also plan to develop tools to support stakeholders, such as moderators, perpetrators, and community members, to collectively make sense of and address harm.

Limitations

We studied SnuggleSense in an experimental setting. Applying these results to commercial-sized platforms and broader use cases would require adapting the design to suit the specific demands and complexities of those contexts.

Our pilot and study participants primarily comprised college students in the United States, potentially limiting the generalizability of our findings to other survivor demograph-

ics. Additionally, the harm scenarios shared by participants may not encompass the full spectrum of online harm experiences. Our initial dataset contains over 200 action items, with survivor similarity calculated based on a limited set of harm experience dimensions. This may limit the diversity of recommendations provided to participants.

Sensemaking is a dynamic and evolving process, influenced by various factors over time [174]. Our study imposed a finite timeframe for participants to make sense of a given harm scenario. It is plausible that participants' perceptions of the same incident might undergo changes with extended time for sensemaking. Furthermore, it is imperative to recognize that sensemaking represents an initial step towards addressing harm. The subsequent action taken is integral to empowerment of survivors [186]. Therefore, a comprehensive assessment of our system's effectiveness can only be achieved through evaluating its impact in the later stages of executing the action plan, which constitutes an avenue for our future research.

5.6 Conclusion

Our paper introduces SnuggleSense, a system designed to empower survivors of online harm by guiding them through a sensemaking process. Inspired by restorative justice, SnuggleSense opens up new opportunities for survivors to assert their agency and define their paths toward healing and resolution. SnuggleSense represents a step forward in empowering survivors of online harm centering their needs and agency in the sensemaking process and highlighting the importance of providing them with the tools, support, and community-based resources to address harm.

Chapter 6

RQ3: What are the opportunities and challenges of implementing restorative justice in the current moderation landscape?

6.1 Introduction

Social media platforms frequently address online interpersonal harm, such as harassment, through content moderation; this involves reviewing user-submitted content for appropriateness and sanctioning contributors that violate the platform's rules [62, 131]. However, despite efforts in research and industry to improve moderation practices in recent years, the number of people experiencing severe forms of harassment continues to grow. In 2014, 15% of Americans reported experiencing severe harassment, including physical threats, stalking, sexual harassment, and sustained harassment [47]. That number grew to 18% in 2017 and 25% in 2021. Further, many people report simultaneously experiencing multiple forms of severe harassment [48, 170]. Research shows that online harms are insufficiently addressed by platforms' and communities' current approaches [49, 36, 105, 173], and 32% of Americans say that social media companies are doing a poor job at addressing online harassment on their platforms [170]. Alternative approaches are desperately needed, but what principles should guide them, and how would they work in practice?

Restorative justice is a framework that argues for repairing harm and restoring individuals and communities after harm has occurred. In this paper, our goal is to *draw from restorative justice philosophy and practice to study how an online gaming community currently addresses—and might alternatively address—interpersonal harm.* We focus on restorative justice here because it has an established offline practice and has been successfully institutionalized to address harm in other contexts, such as schools and prisons [100, 102, 13]. In recent years, the HCI and CSCW communities have also explored its utility in addressing

online harm [20, 145, 95], and we build upon these efforts, as well.

Restorative justice focuses on providing care, support, and in other ways meeting people's needs after harm has occurred. It has three major principles: (1) identify and address the victim's needs related to the harm, (2) support the offender in taking accountability and working to repair the harm, and (3) engage the community in the process to support victims and offenders and heal collectively [114, 184]. In practice, restorative justice addresses harm differently than more common punitive models. The main tool for action in a punitive justice model, as embodied in content moderation, is *punishing* the rule violator. In contrast, in restorative justice it is *communication* among the harmed person, the offender, and the community. For instance, in a common restorative justice practice called a *victim-offender conference*, the victim and offender meet to discuss the harm and how to address it under the guidance of a facilitator; interested community members are also invited to join this conversation since the conference aims to address the needs and obligations of all three parties involved. A follow-up process may include apologies or community service by the offenders [184].

This paper uses the three restorative justice principles described above and its common practices (e.g., the victim-offender conference) as a vehicle to study the perspectives and practices of victims, offenders, and moderators during instances of online harm. First, we use the principles to evaluate current practices for addressing interpersonal harm and identify the potential need for restorative justice practices. We focus on the experiences of victims, offenders, and moderators, who are key stakeholders and participants in restorative justice conferences (with the moderator acting as facilitator). Second, we use the victim-offender conference to identify the benefits and challenges of practicably implementing restorative justice practices in an online setting.

We study harm cases in the Overwatch gaming community, which spans two major platforms : the Overwatch platform on which the game is played [1] and the Discord platform [2] on which gaming discussions, teammate selection, and match organization occur. Online gaming communities have long suffered from severe and frequent incidents of online harm [2, 71, 16]. Our analysis of Overwatch, a multi-player game, lets us explore such harm in the context of different types of user relationships, including competition and collaboration. We interviewed self-identified victims (people who have been harmed), offenders (people who have harmed others), and moderators who dealt with the cases being discussed. Our interview protocol resembles the restorative justice practice of *pre-conferencing*, which is used to learn people's history and preferences, explain restorative justice to them, and evaluate the appropriateness of holding a victim-offender conference [184]. Additionally, given that restorative justice has been chiefly developed through practice[166], we deepened our understanding of its principles and practices by conducting two interviews with offline restorative justice practitioners.

We find that current, punitive online moderation processes do not effectively stop the

[1] https://playoverwatch.com

[2] https://discord.com

perpetuation of harm. First, content moderation is offender-centered and does not address victims' needs, such as receiving support or healing from harm. Though victims may report individual offenders, they continue to receive harm in a community where abuse is prevalent. Second, content moderation directs offenders' attention to the punishment they receive instead of the damage they cause. When punishment is ineffective, as is often the case, there are no alternative ways to hold offenders accountable. Finally, community members with a punitive mindset may further perpetuate harm by not acknowledging the harmed person's experiences or reacting punitively toward perpetrators or victims, particularly when harm cases are complex and layered.

Our findings show that some current moderation practices align with restorative justice, and a few participants have attempted to implement restorative justice practices in their own online communities. Some victims and offenders also expressed needs that align with restorative justice values. However, applying restorative justice online is not straightforward: there are structural, cultural, and resource-related obstacles to implementing a new approach within the existing punitive framework. We elaborate on the potential challenges of implementing restorative justice online and propose ways to design and embed its practices in online communities.

Our work contributes to a growing line of research that applies alternative justice frameworks to address online harm [42, 36, 145, 121, 67, 68, 137]. By evaluating the moderation practice of the Overwatch gaming community through a restorative justice lens and identifying key stakeholders' preferences for the justice framework, our work sheds light on ways to address online harm that go beyond simply maintaining healthy content and working within a perpetrator-oriented model. We highlight how restorative justice has the potential to reduce the continuance of harm and improve community culture in the long run.

6.2 Background

To provide context for our methods and results, we briefly review the two platforms we study, Overwatch and Discord.

Overwatch and Its Moderation Practices

Overwatch is a real-time, team-based video game developed and published by Blizzard Entertainment[3]. It assigns players to two opposing teams of six. Gamers play in the first-person shooter view and can select from over 30 *heroes* with unique skills. They pair up with random players if they enter the game alone, but they can also choose to pair with selected teammates. During each game, players communicate through the built-in voice chat and text chat functions, but some players also use Discord as an alternative. All players are expected to comply with a set of rules laid out in Blizzard's code of conduct [21]. For example, these rules instruct, "You may not use language that could be offensive or vulgar

[3]https://www.blizzard.com

to others," and "We expect our players to treat each other with respect and promote an enjoyable environment."

Blizzard hires commercial content moderators, who are paid company employees, to regulate its games [131]. Though the company shows users a small set of moderation rules in its code of conduct, it is likely that the company has an internal set of more detailed moderation guidelines to help moderators make their decisions [131]. While moderators do not monitor live games, they handle reports from victims by reviewing game replays and take moderation actions if they determine that users have violated platform rules. Typically, offenders receive a voice chat ban or a temporary or permanent account ban. Offenders receive the decision notification, usually without a detailed explanation. Victims who report the incident usually receive a notice that an action has been taken, but they are not told what the action is.

Discord and Its Moderation Practices

Discord is a popular instant messaging platform that is widely used by Overwatch gamers. On Discord, users can create their own communities, called *servers*, that contain both text and voice channels for real-time discussions. At the time of this study, more than 2000 Discord servers were active under the tag "#Overwatch." Overwatch gamers use these servers to discuss the game, find teammates, and organize Overwatch matches.

Discord moderators are volunteer end-users who regulate their communities and screen posts for inappropriate content [83]. Each community creates its own set of moderation rules. Moderators can sanction users by removing their posts, muting them, or banning them either temporarily or permanently. Since moderators and users both have access to public channels in real-time, moderators can actively monitor harm cases on those channels as they occur. They cannot access private channels, but users can report harmful incidents to moderators through private messages. Some communities use automated moderation tools to detect posts containing inappropriate keywords and issue automatic warnings to posters [83]. We contribute to the study of Discord moderation by showing how different stakeholders perceive and engage with cases of online harm.

The Overwatch Gaming Community Spans Overwatch and Discord

Like many online communities, the community we study spans multiple platforms [53]: Overwatch, the gaming platform, and Discord, a discussion platform. Though Overwatch pairs up random players if they enter alone, Overwatch gamers frequently use Discord to discuss the game, stay connected, and communicate with one another. Harm can occur both for players with pre-existing social connections and those who are strangers to one another. Since volunteer moderators participate in the Discord community and are often gamers, they may also be friends with a victim and/or offender in a harm case. In addition, a harm case may initiate in Overwatch but extend to Discord, or vice versa. Our research investigates

harm cases on Overwatch, Discord, or both platforms and includes players with diverse social relationships.

6.3 Methods

In total, we interviewed 23 participants from the Overwatch gaming community for this study (Table 6.1). To understand Overwatch gamers' perspectives on the restorative justice process, we interviewed victims (those harmed), offenders (those who harm)[4], and Discord moderators who dealt with the cases being discussed. Some participants fall into more than one of these three groups. We could not include Overwatch moderators in our study because they are commercial content moderators [132] who remain anonymous and constrained by non-disclosure agreements.

Restorative justice primarily evolved through practice rather than as an academic discipline. To more deeply understand how its principles might be applied online, we conducted two additional expert interviews with facilitators from a restorative justice center at the University of California, Berkeley. Further, the first author attended 30 hours of restorative justice training courses to learn how it is practiced in local communities. These interviews and the training helped ground our research in restorative justice values and practices.

Recruitment

We recruited participants using a combination of convenience sampling and snowball sampling [133, 17]. First, we joined multiple Discord communities focused on Overwatch and reached out to moderators, sending private messages to request an interview. After building rapport with moderators through interviews, we asked their permission to publish recruitment surveys in their communities to find victims and offenders of harm. Some moderators referred us to their fellow moderators for interviews and invited us to other Overwatch Discord communities they were involved in. In total, we recruited participants from five Overwatch Discord 'server' communities. In addition, we recruited two facilitators for expert interviews from a training program the first author participated in. We recruited victims and offenders separately through two surveys. In the survey for victims, we described our

[4] We use the terms "victim" and "offender" for brevity and to clarify participants' roles in specific harm cases. We recognize and agree with calls for eradicating the use of these labels over the long term. Some restorative justice practitioners believe these labels and their meanings are rooted in the punitive justice system, and transformation to restorative justice requires transforming our language. Using "offenders" may imply that people are "inherently bad" and deserve the condemnation of society [24], while using "victims" may feel disempowering to some and deny the agency victims should have in restorative justice [58]. Less popular alternatives to these terms include "the person who caused harm" or "perpetrator" and the "person who has been harmed" or "survivor." Accordingly, the language of "victim-offender conference" has also shifted to "restorative justice circle." However, such alternative terms may also not align with the self-image of harmed participants, as we found during our data collection. For this early-stage research, we retain the original terminology to be consistent with the language participants used.

recruitment criteria as people who have experienced online harm on Overwatch Discord or in an Overwatch game. During the interviews, some victims referred us to their friends who have experienced harm or have been banned in the Overwatch gaming community, and we included them as participants. For the second survey, we did not describe participants as "offenders" or "people who have caused harm" since prior research suggests that people may not want to associate themselves with those categories, especially when there has been no opportunity to discuss what has happened [80, 82]. Therefore, we described the recruitment criteria as people who have been warned or banned on Overwatch Discord or in the game.

In the recruitment surveys, we asked participants to briefly describe a harm case they had experienced. We selected participants from this survey based on the time order of their replies. Additionally, we conducted preliminary data analysis to categorize the types of harm (e.g., on Discord vs Overwatch; between friends/strangers; within the moderation team/between end-users/between end-users and moderators). We then prioritized participants who had experienced different types of harm for interviews. Table 6.1 describes the demographic information of our participants.

Interview Procedure

Through **interviews with victims and offenders**, we wanted to understand both their current experiences with harm cases and their perspectives on a restorative justice process for those cases. We adapted our interview questions from restorative justice pre-conferences, where facilitators meet one-on-one with victims and offenders to solicit their perceptions on using a restorative justice process to address harm [184]. During the first stage of the interviews, we asked participants questions about the harm case they had experienced or caused, including how it was handled, its impact, and the need to address the harm.

During the second stage, we introduced participants to restorative justice principles and the victim-offender conferences. We focused on these conferences because they constitute a widely used practice that embeds the core restorative justice principles. We included frequently posed questions in preparation for victim-offender conferences, such as what information they would like to convey in the conference and their expectations and concerns regarding the process. If time allowed, we asked participants to reflect on more than one harmful incident. Further, some victims and offenders were also moderators of the community. We first inquired about the harm with their primary self-identified roles and then asked about their secondary role(s).

Our **interviews with Discord moderators** were intended to assess how they deal with harm in their communities and their attitudes toward using restorative justice to repair the harm. Additionally, since the facilitator is essential in offline restorative justice practices, we explored the possibility of creating a corresponding role in online scenarios. Since moderators have the closest currently existing role to a facilitator, we sought to learn their perspectives on assuming this role. During interviews with them, we first asked about the harm cases they had handled in their communities and their decision rationales. We then introduced the idea of a victim-offender conference and asked them to (1) reflect on potentially using it

Table 6.1: Participants' demographic information. We recruited participants using surveys for Overwatch or Discord users who (1) have been harmed or (2) have been banned or warned. Additionally, we recruited moderators on Discord. We show here the demographic details of each participant and their self-identified role (victim, offender, moderator, or facilitator; marked by 'x') in the harm cases they discussed with us. Note that a single person may have multiple roles in a harm case or across different cases. We also recruited two restorative justice facilitators.

	Age	**Gender**	**Race/ ethnicity**	**Education**	**Country**	**Victim**	**Offender**	**Moderator**	**Facilitator**
P1	25	Female	Asian	Master's degree	US	x		x	
P2	20	Male	White	Bachelor's degree	UK			x	
P3	24	Non-binary	White	Some college	UK			x	
P4	20s	Female	White	Associate degree	CA			x	
P5	26	Female	Mixed	Master's degree	US				x
P6	27	Male	Fula	Master's degree	US				x
P7	18	—	—	Some college	UK			x	
P8	19	Female	White	Some college	UK	x			
P9	20	Male	White	Some college	US	x			
P10	20	Male	Hispanic	Some college	N/A			x	
P11	21	Female	White	Associate degree	US	x			
P12	18	Female	Mixed	Less than high school	US	x			
P13	19	Female	White	High school graduate	CA	x			
P14	24	Female	White	Bachelor's degree	CA			x	
P15	18	—	White	Less than high school	NA			x	
P16	25	Transgender	White	Master's degree	US	x		x	
P17	18	Male	Asian	Some college	US	x	x		
P18	23	Male	White	Bachelor's degree	UK	x		x	
P19	30	Male	White	Master's degree	UK	x	x	x	
P20	18	Male	Asian	Some college	CA	x	x	x	
P21	18	Male	White	Less than high school	US	x	x		
P22	20	Male	White	N/A	UK	x	x	x	
P23	18	Male	White	High school graduate	Ireland	x			
P24	18	Male	Berbers	Less than high school	Algeria	x	x	x	
P25	24	Female	White	Bachelor's degree	CA	x		x	

Participants' gender and race/ethnicity are self-identified.

as an alternative approach for addressing harm cases they had handled in their communities and (2) share their thoughts about serving as restorative justice facilitators on those cases.

It is challenging to elicit people's perspectives on a hypothetical process or a process they lack previous knowledge about. To make restorative justice concepts more concrete, we asked participants to imagine a restorative justice process based on actual harm cases they had experienced or handled. Additionally, we answered their follow-up questions and corrected any misconceptions we identified in our discussions. We continued analyzing our interview data as we recruited and interviewed more participants. We ceased recruiting when our analysis reached theoretical saturation [30].

We conducted **two expert interviews** with restorative justice facilitators to elicit their insights about using restorative justice in online settings. We introduced these facilitators to Discord and Overwatch moderation mechanisms and described examples of how harm cases were handled based on our interviews with victims and offenders. Here, we stayed close to our raw data and described the harm cases through the perspectives of our participants. The facilitators evaluated the current moderation practices through the restorative justice lens and envisioned the future of restorative justice on Discord and Overwatch. We did not intend to reach theoretical saturation for this population [30]; we still incorporated these interviews because doing so provided valuable insights on how these cases could be alternatively handled in a restorative justice context [138].

We conducted our interviews from February to July 2020. The interviews lasted one to two hours each, and participants received compensation from $25 to $50 US dollars based on interview duration. We conducted 21 interviews using Discord's voice call function, and two participants (P1, P13) chose to be interviewed using Discord's text chat. Before each interview, we negotiated interview time, which was based on participants' availability and the number of harm cases they wanted to share during the interview. We also conducted two in-person interviews with the facilitators (P5, P6). Our study was approved by the Institutional Review Board (IRB) at the University of California, Berkeley.

Data Analysis

We conducted interpretive data analysis on our interview transcripts [30]. We began with a round of initial coding [136], applying short phrases as codes to our data line-by-line to keep the codes close to the data. Examples of first-level codes included "impact of harm," "creating new account," and "banning." Next, we conducted focused coding [136] by identifying frequently occurring themes and forming higher-level descriptions; second-level codes included "notion of justice" and "sense of community." Coding was done iteratively, where the first author frequently moved between interview transcripts and codes and discussed emergent themes with other authors. After these initial and focused coding rounds, we applied restorative justice principles and values as a lens to guide our interpretations. Finally, we established connections between our themes to arrive at our findings, which we categorized according to participants' roles (offenders, victims, moderators, and facilitators). We coded the expert interviews using the same code book we used for other interviews since

it helped us compare facilitators' views of harm to other participants' opinions during the analysis.

Methodological Limitations

We used convenience and snowball sampling to recruit participants in a single gaming community. In addition, our research examines harm from an interpersonal perspective. As a result, the experiences participants report and the solutions we design may not be representative of, or applicable to, all forms of gaming communities or online harm. Further, though Overwatch has a multinational and multicultural user base, most participants were English speakers from Western cultures. Finally, our recruitment method did not let us recruit offenders who were neither warned nor banned.

As researchers, we lack a full picture of what occurred in the harm cases we studied. To this end, we adopted Whitney Phillips's approach [127], which involved observing how our participants presented themselves to us and drawing conclusions from their performance, which may have been choreographed. As a result, we present our findings as subjective perspectives of Overwatch members rather than as objective truths. During the interviews, all victims believed that they were the party that had received harm; all offenders believed they had caused harm, but some assumed only partial accountability since they were also harmed in the process. We analyzed our interviews based on these self-described roles and views of participants in each harm case.

6.4 Findings: Current Moderation Models Through a Restorative Justice Lens

We use the three restorative justice principles (see Section 1) to understand the experiences of victims, offenders, and the community during instances of online harm within the current content moderation landscape. With our sampling methods, we included two victim-offender pairs in our study — P17 and P25 as well as P22 and P24. To illustrate our findings, we present a relevant harm case from our data in each section.

We warn readers that the following sections contain offensive ideas and language. However, we believe that such sensitive content can help readers understand the nature of online harm.

Victims Have Limited Means to Heal from Past Harm or Stop the Continuation of Harm

Restorative justice centers on victim needs. In offline restorative justice processes, victims usually share their stories, receive emotional support, and provide input on what is needed to repair the harm. The process aims to help them heal from harm and stop harm continuation [184]. Our analysis indicates that the main tool available to online victims to address

harm is reporting their offenders to moderators or the moderation system. However, it also shows that reporting such incidents does not effectively address their need for emotional healing or prevent future harm.

To demonstrate the harm participants receive in online gaming communities, we first relate an example in Case 1.

Case 1
P13 teamed up with two previously unknown players to play an Overwatch game. She said *hello* in the voice chat but instantly regretted it: *"These guys started saying 'Is that a girl damn are you cute,' and making jokes."* P13 chose to remain silent, but she soon received a lurid threat: *"They got harsh and said, 'If you don't respond I will take out my dick and slap you."'*

P13 confronted those players, but the players "*all acted like it wasn't a big deal.*" They continued to make jokes about her and complained about her gaming skills after they lost the first round. P13 decided to leave the game but continued monitoring the in-game chat: "*They were complaining about me, saying girls shouldn't play games. It makes me feel so nervous, I started to cry.*"

Several participants told us that they frequently experience harm, such as offensive name-calling or sexual harassment, on Overwatch or Discord. We find that such cases are often related to structural issues such as sexism, misogyny, transphobia, and homophobia—offenders often target their victims' gender or gender identity. Victims also received negative comments about their gaming skills. Women and gender non-conforming gamers sometimes experience compounding harms due to both of these patterns. P11 offered some examples of the comments she had heard while gaming:

> *"'Go make me a sandwich, calling me a bitch, telling me that I should go make food for my husband, I probably have kids and cats [. . .] getting called bitch, slut, whore, cunt, just every derogatory name for women in the book."*

On both Overwatch and Discord, the main tool for victims to address harm is reporting to the moderation system or the moderators. However, victims often felt left out of the decision-making process after reporting a harm case and did not find the process helpful for healing from the harm. Several participants told us that after reporting, they either do not hear back or receive a vague message that indicates the moderation decision but offers no procedural details: *"[The moderation system] just tells me that 'something' happened, and my report led to that happening"* (P11). P9 believes that the moderation system is intentionally opaque to gamers: "They don't want players understand[ing] the system."

In addition, though reporting the harm may result in punishment of the offenders, it does not directly help victims heal from the emotional impact of the harm. P11 talked about the emotional toll of the incident: *"I've cried about it a few times [. . . and I told my friend], 'I don't even feel like I want to live in a world where there are people like that."'*

Many victims indicated that though they tried to report their offenders, they continued to be harmed in the community, where a culture of harm is prevalent. Even if they do not encounter the same offender again, incidents of harm are so frequent that they come across new offenders repeatedly: *"At least 5 out of 10 games, I get somebody saying some sort of crude comment about sexually harassing me"* (P25). When those experiences of harm accumulate, victims anticipate being frequently subjected to harm, and they accept that there are no effective ways to address it. P9 said, *"[Harm] happens frequently enough that you just get used to it."* P11 said, *"I feel pretty helpless that I have to endure that every time I play a game."* P12 noted that reporting itself becomes labor intensive: *"[Harm] happens so often. I wouldn't want to report every single person I've talked to."*

Victims indicated that they need to consciously avoid harm in the community, which impacts their gaming experiences. For example, some participants stopped using voice chat out of fear that it would reveal their gender, even though communication with teammates is important for winning. P13 fears engaging with strangers after having experienced ongoing harmful behavior: *"I will never play a game without a friend I trust, or talk in the game without my friends around because I don't trust that there will be even one person that will defend me when it gets bad."* Some participants also leave a game or Discord community where they have experienced harm.

In sum, these findings show that the intensity and frequency of online harm can substantially impair users' online experiences and cause long-lasting emotional damage. Although the current moderation systems on platforms like Overwatch and Discord offer socio-technical mechanisms like reporting to address cases of online harm, the goals, and the lack of transparency and follow-up inherent in these mechanisms, do little to meet victims' need to heal. In addition, the current reporting mechanisms do not substantially reduce offenses within the gaming community because they do not effectively change the culture: even when participants can avoid the original offender, they continue to be harmed by new ones.

Offenders Are Not Supported in Learning the Impact of Their Actions or Taking Accountability

In a restorative justice view, offenders can address harm by acknowledging their wrongdoing, making amends, and changing their future behaviors. Through practices like victim-offender conferencing, offenders learn the impact of their actions by listening to the victim's side of the story. Afterward, they can repair the harm and learn through actions, such as acknowledging harm (e.g., apologizing), taking anger management courses, or doing community service [184, 40].

As our participants report, current moderation approaches often embody graduated sanctions [96]. That is, moderation teams in Overwatch and Discord tend not to ban offenders permanently after their first offense. Instead, they use more lenient punishment first to give offenders a chance to change their behaviors. Our interviews show that many Discord moderators have elements of a restorative mindset: they want to help offenders learn their

wrongdoing and change their behavior by giving them second chances and providing explanations for their sanctions. Several Discord moderators explained their rationale for graduated sanctions as "giving people a second chance." As moderator P14 noted, *"We don't want [the moderation decision] to be a surprise [. . .] and we will actually really want to encourage them to improve."* Some Discord moderators also provide explanations of their moderation decisions. Such messages typically explain the rules and consequences of breaking them and can be posted by a moderator or a pre-configured bot.

Though some moderators have a restorative mindset, the punitive moderation system and rule-based explanations they rely on may fail to provide learning and support for offenders. They may not effectively stop the perpetuation of harm. We show an example through Case 2.

Case 2

P14, a Discord moderator, described a case where moderators in her community gave an offender multiple chances to reform her behavior, but the offender was reluctant to change.

A transgender woman in P14's community had negative experiences with cisgender men in her life and ranted publicly in the community about her hatred of all men. Though the majority of the community are female and LGBTQ gamers, P14 strives to create an environment that is friendly to its cisgender male members: *"[Cisgender men] are not allowed to say horrible things about the female and trans plus members, [so] in return we expect the same courtesy."*

The moderators warned the offending user several times. P14 reflected, *"One of our mods would go in and be like, 'Hey, just so you know, this is not okay here. I'm going to remove your message and just don't do it again.' She would respond with, 'Oh yeah, got to protect the cishet [ed. cisgender and heterosexual] men from the trans people."'*

This user was temporarily banned after several warnings. She then messaged moderators, telling them that they *"don't know anything about oppression"* (P14). The moderators defended themselves, arguing that they understood oppression, and directed her to the community rules. Subsequently, the moderators stopped engaging with this member.

The temporary ban caused the transgender woman to lose her gender pronoun tag, which demonstrated her gender identity to the community. When she realized this had happened, she swore at the moderators, leading them to permanently ban her from the community.

In case 2, the moderators made efforts to negotiate with the transgender woman, hoping that she would change by giving her multiple chances and providing explanations for their sanctions. However, their explanations mainly emphasize that her actions violated the rules and were prohibited, and the moderation decisions are all punitive. We argue that this model

directs offenders' attention away from the victims and does not further their understanding of the impacts of their actions on the victim. Instead, the punishment, warnings, and restating of rules position the offenders against the moderation system and in defense of their own behavior.

Restorative justice recognizes that often, offenders of harm may have been the victims of harm in other cases. As activist Mariame Kaba says: *"No one enters violence the first time by committing it."* [66]. In case 2, the harm that the transgender woman caused is a reaction to the harm she experienced elsewhere from cisgender men. Similarly, several offenders in our sample revealed how they had been constantly harmed in gaming communities: *"I have been called countless names that are vulgar, offensive, and stuff like that" (P21).* Restorative justice practices assume that although the past harms experienced by offenders do not absolve them of their responsibility for committing an offense, it can be hard to stop offensive behavior without addressing their sense of victimization through support and healing. Punishment, on the contrary, usually reinforces the sense of victimization [184].

The moderators in case 2 stopped the transgender woman from sharing her story. Instead, she received a denial, which worsened the damage and led her to defend herself more aggressively. In addition, when harm such as discrimination is a systemic issue in the gaming community or the broader society, stopping the perpetration of harm may require work to change the culture in addition to work to change individual offenders. Facilitator P6 explained how a restorative justice approach would strive to provide support and look at the root cause instead of punishing a particular offender:

> *"Rather than just zooming in on that one case and trying to blame this individual for that action, acknowledging that [they were] also harmed by the community and pushed to act in that way [. . .] it's a much harder conversation to engage with, but then, that takes away from saying that person is wrong. It's more about [. . .] how do we understand this collectively, and then, how do we address this collectively? "*

We also find that when punishment is used as the only tool to stop offenses, it loses its function when offenders are not actually punished. Several moderators, victims, and offenders mentioned the convenience of using alternative accounts once one account has been moderated: *"It is so easy to make new accounts in Discord that 'reporting' people don't really work"* (P1). P17 used to conduct multiple offenses in games. He pointed out the fallacy of regulating with banning while there is no cost to creating and using another account in the game:

> *"Everyone right now is happy with that system because, if I was a normal player [. . .] I heard that [the offenders] got banned, I would be like, 'Wonderful, they got banned. I'm never going to see them again,' but in actuality, when I got banned I'm going to say, 'I don't give a fuck, I'm just going to log into my second account."*

Despite these shortcomings, we found evidence that the current punitive approach contributed to maintaining the health of community content and stopped the continuation of harm. Moderators we interviewed told us that some offenders would stop their misbehavior after receiving a warning or would not return to the same community after being banned. However, when punishment is the *only* means of addressing harm, it cannot always be effective.

Gaming Communities Create Challenges for Victims and Offenders to Address Harm

Harmful actions disrupt relationships within a community and potentially affect all its members. Thus, it is important for community members to acknowledge the harm and participate in redressing it collectively [184]. Restorative justice defines a *micro-community* as the secondary victims who are affected by the harm and the people who can support the victims and offenders to address the harm (e.g., family or friends) [114]. In online gaming communities, the micro-community includes not only victims, offenders, and moderators but also bystanders, friends of victims and offenders, and gamers, more generally.

However, we found that the relevant stakeholders had no shared sense of community during many instances of online harm. As a result, the harm was often considered "someone else's problem" and remained unaddressed. Additionally, when micro-community members got involved, they often created secondary effects that further harmed victims and offenders.

Victims, offenders, and moderators do not have a shared sense of community

Restorative justice appeals to the mutual obligations and responsibilities of all community members to each other as necessary to address harm. However, in current moderation systems, we found that victims, offenders, and moderators lack a shared sense of community.

Many victims in our sample relate to and show care toward other gamers the offenders might harm. For example, P11 said, "*I usually am just sad not for myself really, but just that other people have to deal with those people.*" Some victims even care about their offenders and what may have led them to perpetrate harm. However, there is a lack of *shared* sense of connection from the offenders, which makes it hard for them to care about their victims and may lead them to fail to see the impact of their actions on others. The anonymous and ephemeral nature of online conversations deters offenders from relating or caring about the community or the victims. P17 used to harm other gamers but has since reformed himself. Reflecting on his previous mindset as an offender, he pointed out: *"[The victims'] day is legitimately ruined because of what [offenders] said, but these people aren't going to think about what they said twice [. . .] They're not going to reflect because it doesn't affect them."*

Additionally, Overwatch randomly pairs up gamers when they do not join as a team. As a result, the offenders do not need to interact with their victims after a game has finished. This absolves offenders from feeling accountable for their actions and in fact creates an environment where repeating harm comes at no cost. P21, who has attacked others in

Overwatch games, noted: *"In a game where you know somebody for 20 minutes and you see that they're bad so you flame [insult] them, and then you just move on. You never see them again."*

As mentioned in Section 5.2, many gamers we interviewed have multiple online identities in games and do not feel particularly attached to one. P17 used to attack others online using his anonymous accounts. He reflected, *"No one can recognize what accounts we're on so we can do whatever we want."* Given the limited information and social cues online, it is harder for people to find mutual relationships and connections. P21 used it to justify offenders' behavior:

> *"In games you don't know these people. You don't know their personalities. You don't know their backstory [...] All you know is that they're doing bad [at gaming], and so that's what you use against them. And, there's no way of fixing that. And if you're getting offended by some of these words, then just mute them."*

On many Discord servers, moderators regulate a confined, public space — the general chat. Several moderators told us they do not intervene in harm cases that happen outside this general chat. Moderator P4 noted: *"For the most part, we just tell people we can't moderate things that happen outside of our community, and at that point it's on them to block people."* For harm cases in the general chat, moderators do not help offenders and victims mitigate problems; instead, they punish the offender or ask both parties to resolve problems themselves. Moderator P19 would move contentious conversations to a private space: *"I'll delete all messages that they've put through to each other, put them both into the group chat, ban them from talking in general or whatever and keep them in this private chat and get it all solved in there."*

In a restorative justice view, a sense of community is essential for offenders to care about the harm they cause to the victims and may even stop them from conducting harm in the first place. However, we found that the anonymous, ephemeral nature of online interactions makes offenders careless about breaking relationships with victims. Additionally, the unclear distinction between public and private online spaces creates challenges for moderators to support victims, and the ensuing lack of support may leave victims vulnerable.

The community creates secondary harm against the victims

In some offline practices of restorative justice conferencing [184], community members who care about the harm that occurred can choose to join the conference. Community members can support victims by listening to their stories, acknowledging the harm that occurred, and providing emotional support [184]. Analyzing our interview data, we found that due to the absence of such a victim-centered structure online, community members often create secondary harm. Though it takes courage for victims to share their stories, finding an audience that supports and cares about them can be difficult. Instead, victims may be challenged, humiliated, or blamed, which creates secondary harm.

P12 was sexually harassed by a male member in the gaming community when she was underage. She chose to reveal the member's behavior during a voice chat when other community members were present: *"I felt like more people should know about it since most of the people in [the Discord community] are underage."* Though P12 expected to get support from her friends, she received multiple challenges from them after this revelation: *"Some people were on my side [. . .] but some other people were on his side and said I just wanted attention, and I should have known he was just joking."* Later, the man denied all accusations and accused P12 instead. According to P12, *"He was trying to make me look like a bad person."* P12 felt so unsupported and hurt by the community's reaction that she eventually chose to leave it.

Community members often ignore the harms they witness, which can further fuel harm. P13 was a victim herself and has witnessed harm in the Overwatch community. She talked about how she was disappointed by the bystanders who did not speak up for the victims:

> *"Most people just think that the game is over; there is no point in trying to help even your teammates, and think they will just end the game and try again another game. Since it is online, most people don't realize that it can actually hurt people what someone says."*

Further, community members can feed into harm by rewarding offenders with attention and positive reactions. P17 talked about why he used to harm intentionally to get attention from friends:

> *"All my friends found it funny. The reason I said those things was not for [the victim . . .] It's for them [my friends] to laugh. It's for the reaction. It's the adrenaline rush of being the center of attention, you know?"*

These findings suggest that community members are not neutral bystanders when harm occurs. They may create secondary harm for victims by ignoring the harm, challenging their stories, or encouraging offenders' behavior. In contrast, restorative justice enables the community to build a culture of responsibility and care, where community members collectively prevent and address harm.

The community has a punitive mindset toward offenders

Restorative justice encourages the community to facilitate change in offenders' behaviors. Community members can share how the harm has impacted them, express willingness to support offenders through the change, and acknowledge any changes offenders have made. However, in punitive settings, community members who disagree with the offenders' behavior want to punish them, as is the case in online moderation systems. Case 3 shows an example of such an occurrence and its effects in Overwatch.

Case 3

P17 and his male friend were gaming as a team with a female player, P25, whom they started to attack with sexist comments after losing the first round. P25 decided to record this incident and posted it on Twitter after the attack continued for a while: *"I knew that it was something that a lot of women and a lot of people deal with daily almost within the community. I wanted to be able to show just how bad it can be."* Many people showed empathy and support for P25, and to date, the tweet has received more than 100 retweets and 600 likes.

After P25 reported P17 and his friend through Overwatch, their accounts were temporarily banned. However, for P17, the account banning was not the most severe punishment; rather, it was the subsequent harassment and damage to his reputation due to P25's public revelation. People located P17's multiple social media accounts. He reflected, *"I would get random messages throughout the day saying, 'You should kill yourself,' and death threats like, 'I'm going to come. I will."'* He abandoned his previous online identity completely: *"I deleted my Twitter, deleted Instagram, deleted Discord, changed all my Overwatch account names."* P17 later apologized to P25 and has since changed his behavior. However, P17 believed the incident changed his career path as a 16-year-old: *"To be honest, I think that was the main reason I didn't try to pursue to go pro in Overwatch harder [. . .] because of how much I had to do to get back the reputation."*

P17 suffered bans and community condemnation, which stopped his offensive online behavior. However, no one supported him in taking accountability and changing after learning the impact of his actions except for one gamer friend who reached out to him. This friend suggested that P17 interact more with people offline because he seemed emotionally detached from his online victims. At that time, P17 used to play video games all day. P17 took that advice, and as time went on, he began realizing the impact of his actions on others: *"[By] talking to real people and interacting with them face to face so you can see their emotions [. . .] now I can imagine them [people I attacked in games] sitting at their computer and just like crying [. . .] and that's why I don't say these things."*

In Case 3, victim P25 had a positive experience when sharing her experiences of harm on Twitter. However, most other gamers decided to punish P17 instead of telling him how his actions caused harm. As we mentioned in Section 6.4, the punishment only further harmed P17: *"It's just anxiety. That's what's constantly going through your head."* This punishment stopped P17 from offending again, but he did not learn to care about his victims until his friend reached out and supported him.

Additionally, though P17 apologized to P25, stopped those behaviors, and learned the

impact of his action afterward, he did not have a chance to demonstrate his change to the community. He left the community and abandoned his career goal to become a professional gamer. This is not what his victim, P25, had wished would happen. She wanted the offender to have a chance to learn and demonstrate his change: *"Nobody's perfect, everybody makes mistakes. We're all human [...] I don't think that just because you did one bad thing a year ago means that you just don't have a chance anymore."* The community response also runs counter to restorative justice views, which maintain that offenders will be welcomed back to the community after their reform [184].

In general, we found that some community members have a sense of responsibility and care about the victims and the harm that occurs inside their community. However, how they address harm is largely shaped by the punitive justice values prevalent in the gaming community and broader society. Their actions of punishing the offenders can create further harm in the community and do not help offenders who may choose to reform themselves.

6.5 Findings: Online Restorative Justice Possibilities and Challenges

We now discuss how the various gaming community stakeholders responded to the idea of using one common restorative justice practice, namely, a victim-offender conference, to address harm in online gaming communities. Our goal was not to measure the participants' binary response regarding their interest in attending the victim-offender conference or use it as a direct measure of the potential of online restorative justice practices. Rather, we sought to identify the essential prerequisites and potential obstacles associated with designing and implementing new forms of online restorative justice by encouraging participants to reflect on their needs and concerns as related in the victim-offender conference.

Victims' Needs and Concerns for a Restorative Justice Process

Many victims wanted to join a victim-offender conference if the offenders were willing to repair the harm. Concurrently, they had concerns and doubts about the process, especially about offenders' readiness to attend the conference. In our sample, no participants mentioned punishing offenders as a desired outcome of the victim-offender conference. We describe these needs and concerns below.

Some victims want to understand and communicate with offenders

Some victims wanted to understand why offenders harmed them and communicate to the offenders how they were hurt. Victims observed that the harm against them occurred unexpectedly, and they could not rationalize the offenders' behavior, which resulted in a need to understand it. P19 said, *"I just want to know what goes through their head at that point in time."* P11 expressed her confusion and frustration:

> *"I just don't understand the motive and don't understand the reasoning. Do they have girlfriends? [...] Or their sisters? [...] Because, I'm a sister, I'm a girlfriend, I have men in my life that I love and they would never do that to me. So why would you do it to someone else's loved one? I don't know."*

Several victims wanted to tell the offenders specifics of how they were hurt. P23 said, *"First and foremost, I would express my feelings, how I felt and how that hurt me."* Some victims hoped their sharing would help offenders realize the impact of their actions. P13 said, *"Maybe hearing how hurt and scared I felt or feel, they would change their perspective."* As someone who had once harmed others, P17 believed that learning how victims feel about their harm was essential to his change: *"If I knew how people felt in games when I made fun of them, I would probably never make fun of them."*

Some victims want an apology, an acknowledgment of mistakes, and a promise of change from the offenders

When we asked victims what they needed to repair the harm, several mentioned that they would like the offenders to realize their mistakes and issue an apology to start their emotional healing. As P13 said, *"Just their realization of what was wrong with a small apology would make me happy."*

Other victims hoped that the offenders would change their behavior and stop offending. For some victims, causing offenders to reform their behavior was paramount:

> *"I wouldn't want him to do anything personally. I just want him to understand what he did wrong and try to fix it so it doesn't happen again with me or another person." (P12).*

> *"I don't really care to see their ranks [in games] drop or anything; I just want them to change." (P13)*

Some victims have concerns about whether offenders are willing to repair the harm

During interviews, several victims were willing to join restorative justice meetings if the offenders were ready to repair the harm. However, they were concerned about whether the offenders they encountered would meet this condition. Several victims had already attempted to reach out to offenders to resolve the issue but were ignored or dismissed by them. These victims did not think the offenders would be open to participating in a restorative justice conference. P22 said, *"I wouldn't be opposed to speaking to him [the offender] again, but I mean, from previous history, it's going to be difficult to speak to him or be able to trust him again because of his actions in the past."* Some victims also questioned whether the offenders would genuinely want to repair the harm even if they consented to join the conference. P21 worried that the conference is a *"get out of jail free card."* He said, *"People could be as*

offensive as they want, then turn around and be like, 'Oh, I'm so sorry. I won't do it again."'

Most victims who shared these concerns did not want to meet with offenders if they could not ensure there was a genuine desire to repair the harm. One exception was P13, who had a strong preference for offenders to hear her voice even if they might be unaccommodating: *"Even if they [offenders] are aggressive, as long as I had someone I trusted there, I would think it would have the best chance at an outcome [for the offender] to hear my voice."* In restorative justice, victims and offenders meet only when both parties are willing to repair the harm. The facilitator also acts as the gatekeeper to ensure that further harm is unlikely before victims and offenders can meet [22].

Some victims believe the harm they experienced is systemic and addressing it requires long-term efforts

While several victims hoped that the victim-offender conference could reform offenders' behavior, others thought the problems they faced had systemic roots that could not be addressed in a single meeting. P16, a transgender woman who was harmed by someone she believed was transphobic, did not want to meet with the offender because she did not believe the meeting could solve the problem:

> *"I mean, because it's being trans, there's just so much systematic oppression to it [...] It just takes time, and it takes education. It takes advocacy [...] I believe he [the offender] is somebody who probably will change if he's given the right education, the right information, but it's going to take time and it's going to take people becoming more and more accepting of trans people. That's not just a fix that's going to be fixed easily through Discord."*

P25, a female gamer who was verbally attacked while playing Overwatch (Case 3), believed the offender's aggressive behavior was likely shaped by more than just what happened in the video game. As a result, she thought the problem could not be solved within gaming communities alone:

> *"It's more probably deeper rooted than just the video game. It probably seeps into family issues to schooling issues to the environment they've grown up in, etc. You can't really help that without doing more and being there in person."*

Some victims want to move past the harm

Some victims noted that the harm had already occurred and wanted to move past it. Several were emotionally exhausted by the harm. P20 already felt disappointed by the offender's reaction when he tried to negotiate: *"I don't think it's worth trying to regain her [the offender] as a friend [...] because things like this can happen again due to the rash behavior."* P16 described her feelings as *"I'm just kind of over it."*

Some victims thought the harm they experienced was trivial and that they were not severely impacted by it, or that they could reduce the emotional harm through their own efforts. As a result, they wanted to move on with their lives. P8 said, *"If it's more personal, then they should maybe apologize. But [. . .] in this case it's not particularly serious."* P9 gave offenders the benefit of the doubt and felt that moving past the event was an easier option for him:

> *"Maybe the people that were talking inappropriately were just having a bad day or something, or maybe they were genuinely not a good person and rude. But either way, it just seems easier for me to move past it by literally moving past it."*

In sum, most victims have needs other than punishing the offenders. However, they want offenders to engage with such needs only when they are sure that offenders genuinely wish to address the harm. In addition, many victims feel that restorative justice approaches might not be sufficient or ideal to address systemic online harm.

Offenders' Interest in a Restorative Justice Process

During the interviews, we simulated pre-conference practices of restorative justice by asking offenders a series of questions to help them reflect on their experiences of harm and discuss their willingness to join a victim-offender conference. We now describe offenders' thoughts on the process.

Offenders may want to repair their relationships with the victims

One offender (P22) wanted to join the victim-offender conference to repair the relationship with the victim (P24). He admitted that he was emotional when committing the harm:

> *"I have since learned to control my temper as I have gotten older. However, it can still be a problem from time to time. I have an extremely competitive and stubborn personality, so when something doesn't exactly go to plan, it can be difficult for myself to accept and get over it."*

P22 wanted to issue an apology to P24. They were organizing an Overwatch tournament together in a Discord community. P22 hoped that the meeting could help him maintain a professional relationship with the victim: *"I would literally just settle for being civil with one another."*

Some offenders prefer the punitive process over restorative justice

Several offenders (e.g., P17, P21) agreed that they should take full accountability for the harm they had caused. However, they preferred to receive punishment for their actions rather

than join a victim-offender conference. In Case 3, P17 was banned for verbally attacking P25 and lost his reputation in the Overwatch community after P25 posted the video footage online. Though he acknowledged his mistake and apologized to P25 privately after the offense, he did not think he would have attended a victim-offender conference at that time. He described the process as *"boring"*:

> *"I'm not going to go in with an open mind so nothing will get done anyways [. . .] Would an immature 16 year old teenager who's rebellious against his mom, he doesn't want to do the freaking dishes, do you think he'll want to sit in a call or a meeting with the person that he just harassed for the last 20 minutes and figure it out? My answer is no." (P17)*

We found that these offenders' notion of justice aligns with the tenets of punitive justice. They believed they deserved punishment and did not trust the restorative justice process to help them achieve a better outcome. P17 said, *"I definitely fucked up a little bit, so I deserve the punishment."* P21 was banned for insulting others in a game. He was aware of his wrong-doing and expected to get punished even if he joined the meeting, so he wanted to go through the punishment directly: *"If you know for a fact you're in the wrong, then there's no point in even talking, because you're going to get banned anyway."*

Both P17 and P21 were under-aged when they committed the harm. Facilitator P5, who works in a middle school, noted that the resistance P17 and P21 expressed is common in offline practices. P5 argued that the end goal of restorative justice is not to punish offenders but to help them take accountability for their actions and change future behavior. The pre-conference is a chance for facilitators to introduce restorative justice and talk about how it may benefit the offender. In school settings, P5 noted that the support for restorative justice from community members, including parents and teachers, encourages teenagers to be more open to the process:

> *"There's parents' support [. . . and] there's so much research around how restorative justice works in a school and is incredibly beneficial for the students; it's an incredible healing process for the entire school community that a lot of [school staff] are willing to buy in." (P5)*

Some offenders do not fully acknowledge their role in the harm

Several offenders wanted to use victim meetings as a chance to justify their behavior. Those offenders believed that their victims behaved improperly or had hurt their interests in the first place. For example, P20 argued that the supposed victim lied about the situation, and he was wrongfully banned for defending himself. He hoped the victim could come to the meeting without preparation so he could challenge her by surprise: *"They'll present the evidence right on the spot so that she doesn't have time to think or lie or really erase any evidence."* This group of offenders agreed that they should take partial accountability for

the harm they caused. However, they primarily wanted to use the meeting to hold the other party accountable and/or alleviate their own punishment:

> *"Why is it I'm being reported and banned after saying one thing, but yet you're not getting any punishments [for saying many]?"(P19)*

As noted in Sec. 5.2, the current moderation systems lacks a means to hold offenders who use multiple anonymous accounts accountable for their actions. When offenders can conduct harm without receiving any consequence that they care about, participation in restorative justice becomes additional labor instead of an alternative to receiving punishment. As P21 said, *"People aren't going to apologize. This isn't the real world [...] If you get banned, they'll just go play another game."*

In sum, we found one offender in our entire sample who wanted to repair his relationship with the victim, but the mindset of most offenders aligns with punitive justice. Offenders who acknowledge their wrongdoings think that they deserve punishment and do not believe that restorative justice can lead to a better outcome. Those who do not fully acknowledge their wrongdoings want to appropriate the victim-offender conference into a punitive process that holds victims accountable. These views reflect offenders' emphasis on the consequences of their actions on themselves rather than on harm reparation.

Moderators' Views on Implementing Restorative Justice Process in Their Communities

We next discuss Discord moderators' attitudes toward implementing a restorative justice process in their communities. Most moderators agreed with the values of restorative justice, and some of their moderation practices overlapped with its practices. At the same time, they expressed concerns about adapting the moderation process to the restorative justice model. Additionally, several moderators had already attempted restorative justice in their communities but received push-back and challenges from other moderators and community members.

There are elements of restorative justice in the current moderation practice on Discord, but for different purposes

We find that both moderators and facilitators talk with victims and offenders when handling harm cases, but the issues addressed and the end goals differ. In offline restorative justice practices, facilitators talk with victims or offenders in pre-conferences, where they ask questions to determine what is needed to repair the harm [128]. Some Discord moderators also speak to victims and offenders before making a moderation decision, but their goal is to make informed decisions on how to punish offenders.

A restorative justice pre-conference happens after fact-finding and focuses on emotions, impact, and the need to repair harm, and the facilitator shows support and empathy throughout the process [167, 22]. On the other hand, the conversation by moderators focuses on

facts and evidence, with the moderators acting as judges. As moderator P3 described, *"We will go and speak to whoever was reporting them [offenders] and speak to people who were involved, and try and get a feel for what actually happened and make a call from there."*

In addition to the process of talking with victims and offenders, both restorative justice and the current moderation system aim to understand offender behaviors beyond the current harm case of interest. Restorative justice situates offenders in their life stories. As noted in Case 2, one life story would be that the transgender woman who offended cisgender men in the Discord community had had negative experiences with cisgender men in her life. A life story can help offenders find their triggers of harm, and the community could then provide support to help them heal [89]. On the other hand, Discord moderators keep logs of past offenses for all users in the community. When an offender commits harm, the moderators review the logs as a reference to determine the proper punishment. As Discord moderator P2 explained, *"We keep logs of all moderation actions that have been taken against any individuals. And so we always check those before handing out any issues [moderation decisions]."* Because of the graduated sanctions mechanism, the punishment is often heavier for offenders with past offenses.

Moderators' power may hinder restorative justice process

Though we may think that the moderators' role is closest to that of facilitators, we find that the power moderators hold may impede restorative justice process. Moderator P7 has a work background in restorative justice, and he tried to implement pre-conferencing with offenders in his Discord community. He believed that he failed to reach desired outcomes with offenders because of his position of power. P7 banned an offender for cheating about his game rank to win. He conducted a pre-conference with him, where he wanted the offender to learn the impact of his actions on the victims he cheated on. However, the offender expressed the wish to get the ban revoked by offering professional Overwatch courses to P7 as an Overwatch coach. P7 was disturbed by the answer: *"It was concerning, right? Because he's answering in a way to try and please me."* P7 believed that the removal of power from online facilitators is essential for authentic sharing:

> *"I don't want that kind of attitude of when people go into a performance they think of like, 'Oh, I have to do well now, because they [moderators] have the decision-making power to remove me from [the Discord server]."'*

Facilitators also pointed out that while moderators have the final right to interpret what has happened and who is right or wrong, restorative justice practices seek to give agency to victims and offenders. Moderators' power may create prejudices against victims or offenders and reduce their agency in the restorative justice process:

> *"As facilitators I'm never like, 'Oh I take side with this story.' I'm always multi-partial, I hold the stories and then it's up for the individuals to figure out what's right to move forward." (P5)*

Though removing punitive power from facilitators may be important, several moderators showed reluctance to let go of power. They worried that giving users agency may lead to unfair and biased outcomes because users may pursue an outcome that aligns with their own interests:

> *"They [the victims] are going to use it as kind of a tool to punish people that they don't like." (P25)*

> *"[The community members in the conference] may initiate a witch hunt or just try and protect their friend group." (P3)*

These moderators' concerns are valid: it requires labor and skill from facilitators to address those potential issues. In offline practices, the facilitator is an essential role that maintains a power balance between victims and offenders, for instance, by ensuring they have equal opportunities to express their opinions. [128].

Some moderators think the labor of restorative justice is disproportional to its gains

The pre-conference and conference processes require significant labor from facilitators [128]. Though a facilitator is usually paid in offline settings such as schools and prisons [89, 86], Discord moderators are volunteers. Several moderators expressed concerns about the labor required to implement restorative justice. Moderator P7 said, *"A lot of [moderators] are volunteers and so the easy option is to just mute people or temp ban people or permanently ban people."* Being a facilitator also requires knowledge of the restorative justice practice. P3 thinks they would need to receive additional training to become a facilitator: *"We're not qualified for this [...] I don't know how we could provide support, or how to make sure that if we give the support, it's beneficial to them."*

We found that many moderators are devoted to maintaining a healthy community environment, and some spend hours handling a single harm case. Moderator P14 gave an example: *"To ban somebody, we actually can have about five or six hours worth of meetings [...] to make sure that our punishment fits the crime essentially."* They were concerned with the restorative justice process because they were unsure whether the extra labor would make any changes. Several moderators believed that users who re-offend multiple times have malicious intentions, so it is not worth spending more effort on them and helping them change. P1 said, *"[Those harms] aimed at our identity (women, queer, trans, etc.) are often by people who enjoy calling us names. So, I don't really see the point of giving these people more opportunities to be prejudiced bigots."* Similarly, P20 thinks offenders *"can't be as civil"* in an online restorative justice conference compared to offline: *"I do agree that having a talk and communicating would be nicer, but [...] it generally doesn't go as well as in real life."*

Some moderators experienced challenges moving from individual restorative justice practices to institutional buy-in

Several participants in our sample had an education or work background related to restorative justice and had attempted restorative justice practices in their gaming community. As noted in Sec. 6.5, P7 conducted a pre-conference with an offender on the Discord server he moderates. P25, an Overwatch gamer, facilitated one victim-offender conference with two friends who fought during an Overwatch game. In the conference, the two friends acknowledged the impact and apologized to each other.

P7 and P25 independently initiated the restorative justice process. However, P7 believed in the importance of engaging the moderation team and other community members to practice restorative justice: *"It takes more than one person to effectively execute restorative justice. For me, I have had a lot of roadblocks when it comes to trying to implement the system [alone]."* In offline scenarios, buy-in at the institutional level (e.g., schools, neighborhoods, workplaces) is important. Institutions can officially establish restorative justice as an alternative to the established punitive justice system [85] and have resources to hire facilitators and remove or mitigate punishment for offenders who successfully pursue the restorative justice process [89]. Community members also gradually familiarize themselves with restorative justice and support victims and offenders in the process [89, 14].

It is difficult for an online community to endorse restorative justice when the established culture and systems are punitive, and no examples demonstrate its effectiveness. As a moderator, P7 promoted restorative justice in his moderation team but failed: *"You've got a traditional model and there's no real examples to demonstrate the capabilities of this [restorative justice]."* P18, a head moderator in P7's team, reflected on people's reactions to P7: *"He (P7) kind of just mentions it [restorative justice], tries to explain it and then everyone gets confused and they kind of step back."* Similarly, P8 tried to promote an alternative moderation system in Overwatch but could not get support from the gamers and Overwatch staff she talked to. She shared how people responded: *"[People said that] the current system worked. It wasn't perfect, but it worked, and our system was new and untested."* Despite such impressions, it is important to consider how we evaluate models for responding to interpersonal harm. For instance, have the victims' needs been met? Was harm repeated? Do offenders recognize the impact of their actions?

In our interviews, we found it hard for people to imagine alternatives to the current moderation system. When we asked participants about their needs beyond having offenders banned or we introduced restorative justice to them, they often found it difficult to imagine the alternatives:

> *"It's not something I've thought about before." (P2)*

> *"I'm not really sure what else you could do. Banned is the thing that everybody's always done." (P9)*

Table 6.2: Comparison of punitive content moderation with an approach based on restorative justice.

	Punitive content moderation approach	**Restorative justice approach**
Victims	Victims are left out of the moderation process.	Addresses the victim's needs for reparation of harm, such as support and healing.
Offenders	Offenders receive rule-centered moderation explanations and punishment.	Encourages offenders to learn the impact of their actions and take responsibility to make reparations.
Community members	Community members may further harm or punish victims and offenders.	Engages community members in supporting victims and offenders and heal collectively.

In sum, many moderators spend time implementing practices that, on the surface, resemble restorative justice (e.g., talking with victims and offenders) but serve punitive purposes. In addition, the prevalence of intentional harm and the wide adoption of punitive models create challenges for communities that want to adopt a restorative justice approach.

6.6 Discussion

As noted throughout the paper, current content moderation systems predominantly address online harm using a punitive approach. Analyzing through an alternate lens of restorative justice values, we found cases where this approach does not effectively meet the needs of victims or offenders of harm and can even further perpetuate the harm. Restorative justice provides a set of principles and practices that have the potential to address these issues. Table 6.2 compares our sample's punitive content moderation approach with a restorative justice one. As the table shows, the latter provides alternative ways to achieve justice from the perspectives of victims, offenders, and community members, who are often absent in current content moderation.

Applying restorative justice practices to the governance of online social spaces is not straightforward. Victims, offenders, and moderators currently have unmet needs and concerns about the process. In this section, we first discuss the challenges of implementing

restorative justice. We then offer possible ways for interested online communities to implement the approach. Finally, we reflect on the relationship between punitive content moderation approaches and restorative justice.

Challenges in Implementing Restorative Justice Online

We now explore the potential structural, cultural, and resource-related challenges of implementing restorative justice online. We also discuss some possible ways to address them.

Labor of restorative justice

Restorative justice practices and the process of shifting to them will likely require significant labor from online communities. Restorative justice conferencing can be time intensive for all stakeholders involved. Before victims and offenders can meet, facilitators must negotiate with each party in pre-conferences, sometimes through multiple rounds of discussion. The collective meeting involves a sequence of procedures, including sharing, negotiating, and reaching a consensus.

Stakeholders, in particular facilitators, must expend emotional as well as cognitive labor. In offline practices, facilitators are usually paid by host organizations such as schools [59]. However, the voluntary nature of moderation on social media sites like Discord means that online facilitators may be asked to do additional unpaid work. This issue can be particularly salient when moderators already expend extensive labor with a growing community [131, 46, 155]. Labor is also involved in training the facilitators. Unlike punitive justice, restorative justice is not currently a societal norm that people can experience and learn about on a daily basis. Aspiring facilitators need additional training to learn restorative justice principles and practices and implement them successfully to avoid creating additional harm.

We estimate that resources for addressing the aforementioned labor needs could be attained in both a top-down and a bottom-up fashion. A top-down process could require resources from companies that host online communities. There is precedent for platforms making such investments; in recent years, social media companies such as Discord have hosted training for its community moderators [5]. A bottom-up process could engage users with preexisting knowledge about restorative justice to first introduce the process in their communities and gradually expand the restorative culture and practice from there. In our sample, two moderators attempted or practiced online restorative justice within their own communities; they showed enthusiasm for expanding its practice to other online gaming communities. It is possible that resources from companies *and* practitioners could collectively begin to address the labor problem.

Additionally, implementing online restorative justice requires *reallocationing* labor instead of merely adding labor. We found that many moderators we interviewed already practiced different elements of restorative justice. Some aim to support victims and give

[5]Discord Moderator Academy. https://discord.com/moderation

offenders a second chance but do not have the proper tools or procedures to achieve that. Other moderators have practices that embed elements of restorative justice, such as talking with offenders and victims. Rather than necessarily requiring new procedures, restorative justice requires a shift of purpose in existing processes – from determining the point of offense to caring for victims and offenders.

Importantly, if online restorative justice could stop the perpetuation of harm more frequently than punitive justice, it could *reduce* the need for moderation labor in the long term. While research has shown that offline restorative justice has successfully reduced re-offense rates [102, 119, 40], evaluating the effectiveness of restorative justice practices in online communities is an important area for future work.

Individuals' understanding of justice aligns with the existing punitive justice model

Although people have needs that the current system does not address, we found that their mindsets and understanding of potential options often align with what the current system can already achieve through its punitive approach. As our research shows, many moderators and victims think that punishing offenders is the most or best they can do, and some offenders also expect to receive punishment. Community members also further perpetuate the punishment paradigm. This mindset is not only a result of the gaming community's culture; it is pervasive throughout society, including in prisons, workplaces, and schools [54, 94].

Given the lack of sufficient information about and experiences with restorative justice, people may misunderstand and mistrust this new justice model. Some offenders in our interviews still expected to receive bans after the process, but restorative justice usually serves as an alternative to punishment, not an addition to it. Some participants wanted to implement alternative justice models in their own communities but received resistance from users who argued that the current moderation system works for them while disregarding its limitations for others. We found that such perspectives usually lead to quick rejections of the notion of implementing restorative justice online.

Before people can imagine and implement alternative justice frameworks to address harm, they must be aware of them. Crucial steps in this direction are information and education. Helping people understand the diversity of restorative justice processes and how their aim is restoration instead of punishment may address their doubts and open more opportunities for change. This is especially important since an incomplete understanding of restorative justice may cause harm in its implementation. For example, enforcing "successful outcomes" may disempower victims and result in disingenuous responses from offenders. Adapting restorative justice to online communities may require changes in the format and procedure of how harm is handled, but prioritizing its core values should help avoid additional unintentional harm.

Restorative justice has developed and rapidly evolved in worldwide practice [184, 39]. Future research can build on and expand restorative justice beyond the three principles and the victim-offender conference. In addition, people need to experience it to understand it,

adapt it to their needs, and learn about its effectiveness. By experiencing it offline, several participants in our sample came to see it as a natural tool adaptable to online communities.

In future work, we plan to collaborate with online communities to implement and test restorative justice practices. We want to pilot online restorative justice sessions and run survey studies to understand the types of harm and types of processes most likely to benefit from these practices. Building on that research, we aim to provide more precise empirical guidelines about how restorative justice can be embedded in moderation systems based on the socio-technical affordances of various online communities. To manifest a future of restorative justice, *"We will figure it out by working to get there."* [87].

While this research focuses on the restorative justice approach, it is not the only alternative and has its limitations. As some of our participants mentioned, many harms are rooted in structural issues such as sexism and racism. *Transformative justice* [38] and *community accountability* [118] are frameworks of justice developed to address such issues. *Procedural justice* emphasizes the importance of a fair moderation decision-making process [73] and is a key objective in restorative justice practices [10]. Future work should explore the potential of these different justice frameworks to address online harm.

Offender accountability

Another challenge may be in motivating offenders to take accountability for their wrongdoing – a persistent moderation problem regardless of the justice model implemented. In our interviews, we found that given the finite scope of moderation in many contexts and the limits in technical affordances of online communities, offenders can often easily avoid punishment. The harm may happen in a place without moderation or clear rules of moderation, e.g., when harm occurs during a private Discord chat or across multiple platforms. Some participants also noted that having multiple identities/accounts in Overwatch or Discord is easy. Thus, when punishment is ineffective, punitive justice may also lose its effectiveness.

Punishment is not the only form of holding offenders accountable. Restorative justice believes that people are also connected through relationships. Our interview data, as well as restorative justice literature, suggest the importance of a sense of community. If offenders perceive themselves as members of a shared community with victims, they will be more likely to take accountability for addressing harm [128, 89]. However, in our interviews, we find that there exists a lack of sense of community. Offenders may not expect to meet victims again or hide behind multiple anonymous accounts. Moderators typically moderate a confined space of general chat, which can leave harm unaddressed in any place outside.

Therefore, it is vital that we inquire into what accountability means to the community and how to hold people accountable within the current moderation system. If offenders can simply avoid any punitive consequences of conducting harm and do not feel a sense of belonging to the community where harm occurs, it would be challenging to engage them in any process –punitive or restorative– that holds them accountable.

Emotion sharing and communication in restorative justice

The limited modes of interaction and the often anonymous member participation in online platforms may influence the effectiveness of restorative justice. Many online interactions are restricted to text or voice, prohibiting victims and offenders from sharing emotions, and may give rise to disingenuous participation. Emotional engagement by victims and offenders is essential for empathy and remorse [156]. Face-to-face sharing lets victims and offenders see each other's body language and facial expressions.

Implementing offline restorative justice includes a series of physical structures and meeting procedures to elicit genuine, equal sharing. For example, participants sit in a circle; individuals who speak hold a talking stick; there are rituals at the beginning of the conference to build connections among participants and mark the circle as a unique space for change [128]. Those rituals for emotion sharing are hard to replicate in the online space. For example, if an offender messages an apology through text, it can be harder to discern a genuine apology from a disingenuous one.

The issue of computer-mediated communication and emotion-sharing has been long discussed in the HCI and CSCW literature. In recent years, increasingly more advanced technologies have been developed to facilitate civic engagement and communication in online systems. For example, Kriplean et al. built a platform, ConsiderIt, to support reflective interpersonal deliberation [97]. REASON (Rapid Evidence Aggregation Supporting Optimal Negotiation) is a Java applet developed to improve information pooling and arrive at consensus decisions [78]. Many researchers have attempted to model human-like characteristics and emotional awareness in chatbots [1, 154].

In the context of the restorative justice approach, Hughes has developed a tool, Keeper, for implementing online restorative justice circles [74]. Such existing systems can be leveraged to develop advanced tools that facilitate emotion-sharing and communication in online restorative justice processes. Online platforms such as Overwatch and Discord could add such tools to improve emotional sharing and communication, necessary conditions for implementing restorative justice.

Applying Restorative Justice in Online Moderation

We now discuss possible ways to adapt current moderation practices to implement restorative justice online. While we have used victim-offender conferencing as a vehicle to interrogate the opportunities and challenges of implementation, restorative justice includes a set of values and practices that extend beyond the conference. It is important to meet online communities and platforms where they are and design restorative justice mechanisms based on their resources, culture, and socio-technical affordances. For example, compared to Overwatch, Discord provides more flexibility in communication and ways of maintaining social connection after a game. Some Discord servers have a greater sense of community and moderation resources than others. We suggest that online communities or platforms can begin with partial restorative justice practices that involve only a few stakeholders (e.g., pre-conferencing)

or adapt some moderation practices to embed restorative justice values (e.g., moderation explanations) and implement victim-offender conferencing when its preconditions are met.

Embed restorative justice language in moderation explanations

Moderators can embed restorative justice language in explanations of moderation decisions. Prior work has shown that providing such explanations can reduce re-offense in the same community. However, many moderated offenders dismiss such explanations and continue re-offending [81, 79]. Our research shows that explanations often focus on describing community guidelines or presenting community norms to the users, not encouraging offenders to reflect on the impact of their actions. These explanations usually indicate the actual or potential punitive consequence, which may direct offenders' attention to the moderation system instead of their victims.

We suggest a shift in language from a rule-based explanation to an explanation that highlights the possible impact offenders may cause to victims and supports the offender in taking accountability. Facilitator P5 provided an example of how she would communicate with the offenders if she were an online facilitator: *"This post had this emotional impact [. . .] this is how you've caused harm. This is the feedback from the community, and we want to work with you to create change."*

Victims are often left out of the moderation process in the online communities we studied. However, some victims want information on the moderation process being used and results of moderator interventions. While prior research has discussed how moderation transparency prevents offenders from re-offending [79], our work highlights the importance of providing information for victims in the moderation process since they are the ones with needs for support and healing. We suggest that a note of care may help victims feel heard and validated and help them recover.

Restorative justice conferences with victims or offenders

In our interviews, we found that some victims or offenders may not be available or willing to meet and have a conversation to address the harm. When a collective conference is not possible, it is possible to apply a *partial restorative justice process* that includes offenders or victims alone [100, 114]. Some Discord moderators already talk with victims and offenders before making moderation decisions, a practice similar to restorative justice pre-conferencing. In offline pre-conferencing, facilitators ask questions to help victims and offenders reflect on the harm, providing them with opportunities for healing and learning.

We propose that pre-conferencing provides opportunities to meet some of the needs our participants identified. For example, when an offender wants to apologize to the victim, the victim may not want to meet the offender but communicate their feelings through the facilitator.

In offline restorative justice, pre-conferencing is a preparation step for a potential victim-offender conference. Thus, if both victim and offender are willing to meet with each other

in the pre-conference, the moderators can organize a victim-offender conference, where both parties share their perspectives and collectively decide how to address the harm. Moderators have the responsibility to maintain a safe space for sharing, for example, to halt the process when harm seems likely to happen, ensure a power balance between the victim and the offender, and work with participants to establish agreements of behavior and values to adhere to throughout the process [22]. In cases where restorative justice does not succeed, preventing the continuation of harm must be prioritized. Because facilitating these processes is difficult, we discuss the challenges and opportunities for moderators to facilitate in the following section.

A victim-support group is another form of restorative justice conferencing [43]. In our sample, many victims indicated a need to receive emotional support. We propose that online communities offer a space for victims to share the harm they experienced and receive support. Systems with similar goals have previously been built in online spaces. For example, Hollaback is a platform that allows victims of street harassment to empower each other through storytelling [44]. Heartmob enables victims to describe their experiences of harm and solicit advice and emotional support from volunteer Heartmobers [19]. These platforms offer templates for how victim support systems can be built. However, support in these platforms is distant from where harm occurs, and it is also important to think about how online communities can support victims by motivating community members to provide support.

Situating Restorative Justice in the Moderation Landscape

We have illustrated possible ways to implement restorative justice in the current moderation system. Yet, we do not seek to advocate for restorative justice as a wholesale replacement of the current moderation approach. We propose that *restorative justice goals and practices be embedded as part of an overall governance structure in online spaces.*

Restorative justice conferencing should be used in select cases only because it is effective only when it is voluntary and the parties involved are committed to engaging. Our findings show that individuals have different conceptualizations of justice. Schoenebeck et al. also found that people's preferences for punitive or restorative outcomes vary with their identity or social media behaviors [145]. It is thus important to attend to victims' and offenders' preferences in individual harm cases.

A larger governance structure should also take into account what to do if restorative justice processes fail. For instance, if it is determined at the pre-conference stage that a restorative justice approach cannot be applied, actions such as removing access by muting or banning might be used to prevent the continuation of harm. This is consistent with offline restorative justice practices, where the community or court clearly defines the types of cases that go through a restorative justice process and the action to take if a restorative justice process is not possible [59, 167].

We estimate that whether an online community decides to apply restorative justice is a value-related question. While restorative justice and punitive justice processes share a goal of addressing harm and preventing its perpetuation, they have different processes and

orientations toward when justice is achieved. Content moderation—closer to a punitive justice approach—addresses harm by limiting the types of content that remain on the site and punishing offenders in proportion to their offense. In contrast, restorative justice aims to assure that victims' needs are met, and offenders learn, repair harm, and stop offending in the future. Thus, the primary reason for applying restorative justice in moderation is not to achieve the current goals of effectively removing inappropriate content and punishing offenders but to benefit online communities using restorative justice values and goals.

Seering et al. found that community moderators have diverse values regarding what moderation should achieve [149]. While some have a more punitive mindset and hope to be a "governor," others align more with restorative justice values and hope to be "facilitator" or "gardener." Thus, online communities must reflect on their values and goals and decide on what mechanisms (e.g., punitive or restorative) help realize those values. Recent research has argued that social media platforms are responsible for incorporating ethical values in moderation instead of merely optimizing to achieve commercial goals [70]. In particular, some researchers have proposed values and goals that align with restorative justice, such as centering victims' needs in addressing harm [145, 19], democracy [148], and education [176]. Our work adds to this line of research and envisions how restorative justice may benefit online communities in addressing some severe forms of online harm, such as harassment.

Finally, communities should be cautious about expecting or enforcing a positive outcome. Enforcing forgiveness from victims or expecting a change in offenders' behavior may undermine victims' needs and put them in a vulnerable place for forgiveness or induce a disingenuous response from offenders [12]. Online communities should allow for partial success or no success without enforcing the ideal outcome, especially at the early stage of implementation when there are insufficient resources or commitments. Instead, they may focus on how victims, offenders, and the entire community could holistically benefit from the process.

6.7 Conclusion

In this research, we interviewed victims, offenders, and moderators in the Overwatch gaming community. We presented case studies that identified opportunities and challenges for using restorative justice in addressing online harm and discussed possible ways to embed a restorative justice approach in the current content moderation landscape. Much remains to be done to explore what restorative justice may look like in an online community context and when and how to implement this approach. We hope this work offers a valuable guide for designers, volunteers, activists, and other scholars to experiment with restorative justice and related approaches in online communities.

Chapter 7

Discussion and Conclusion

In this chapter, I discuss the implications of my dissertation research. First, I talk about how my research contributes to addressing online interpersonal harm. Second, I talk about lessons learned from applying offline justice frameworks in the context of online harm. I conclude my dissertation by arguing for diverse justice and values in the online space.

7.1 A Culture of Reparation: Shifting Focus from Rules to Harm

My dissertation research explores how restorative justice can be applied in the current moderation landscape. At its center, my research argues for a shift from a rule-centric approach to prioritizing harm and its repair. This change in perspective removes barriers inherent in punitive justice frameworks and broadens the methods and resources available for harm mitigation. Below, I detail how this shift in focus opens new pathways for addressing online interpersonal harm.

A multi-stakeholder approach

Shifting our focus from rules to harm and repair enables a multi-stakeholder approach to addressing online harm. Content moderation, with its punitive core, puts addressing harm as an issue between the platforms that issue the rules and the perpetrators who break the rules [62]. In contrast, centering on harm and repairing allows us to begin with the question of the impact of event, needs of affected communities and the stakeholders who are obligated to address these needs.

With this multi-stakeholder viewpoint, my research uncovered a diverse set of stakeholders that can facilitate online harm mitigation, what they can do, as well as survivors' preferences for a timeline of actions from them (Chapter 4). My research also interrogates the preferences of key stakeholders (e.g., survivors, perpetrators, and moderators) in engaging

in this process, and uncovered the opportunities and challenges in the current moderation landscape (Chapter 6).

The central stakeholder in my research is survivors. While their roles are key to reparation, their voices are often overlooked with a punitive-centered approach. My research develops tools and processes for survivors to make sense of the harm and uncover their needs and the available resources (Chapter 4, 5). This process unveiled an arrange of needs from survivors that are unmet by content moderation. These needs open up new opportunities to design social media platforms' policies, features, and harm resolutions. Importantly, actions to address these needs often span through a timeline, with the end being transforming the platform environment. It points to the importance of building a resilient community culture to prevent harm from happening in the long run instead of viewing harm as discrete incidents ruled by moderation.

A cross-platform approach

Focusing on harm instead of rules also allows us to imagine cross-platform processes for addressing online harm. Content moderation is platform-specific, with its rules having boundaries for each platform's public space [62]. A harm-based perspective, instead, transcends these boundaries, acknowledging that harm can span multiple platforms and sometimes connect to offline spaces.

My research points to the gap between the often cross-platform nature of online harm and the limited scope of content moderation. When harm happens on multiple platforms, or both online and offline, it can become "someone else's problem", leaving a grey area that no one is responsible for (Chapter 6). It urges online platforms to rethink the scope of content moderation and their responsibility to ensure users' safety. In addition, it is important to establish accountability from a multi-stakeholder perspective, where bystanders, family and friends, or other survivors fill in the gap that content moderation doesn't reach.

Rethinking the definition and assessment of online safety

Prioritizing harm over rules prompts us to reevaluate our understanding of online safety. For online interpersonal harm, platforms often signal efforts to ensure safety through content moderation efforts such as banning users or removing inappropriate content [62]. In addition, platforms assess the health of the community through tracking the number of moderation actions issued and assess the success of reducing harm through reoffending rates of perpetrators [144]. However, my research suggests that there harm persists even after successful moderation interventions, and safety needs cannot be addressed by content moderation alone (Chapter 4, 5, 6). Survivors still feel unsafe after their perpetrators are removed if the community culture is toxic and they will still constantly receive harm from other perpetrators. Many survivors limit their ways of participation to hide their identity or only play with friends to prevent unexpected harassment. Banning perpetrators also doesn't sufficiently address survivors' needs for emotional safety. In addition, many survivors are

hesitant to report harm given its limited effect in stopping the continuation of harm, thus there are harm not covered by any moderation metrics when the survivors do not resort to this system. With the limited online accountability, some perpetrators are also unimpacted by moderation decisions.

My research suggests that we need to redefine what it means to stop the continuation of harm and to envision online safety measures beyond content moderation. Importantly, we should understand safety from the users' perspectives: Do users feel safe participating in online communities? Platforms need to develop ways to include users' voices in evaluating safety measures and prevention strategies.

Towards a user-centered harm taxonomy

Finally, a harm-focused approach enriches our understanding of online harm as a concept. Most research on online interpersonal harm uses descriptions such as sexual harassment and discrimination to reflect the nature of harm (e.g., [99, 124, 56]). Researchers and industry practitioners have also developed a variety of harm taxonomies to characterize harm on the internet [152, 142, 185]. However, few of these categorizations of harm are directly in alignment with what matters to survivors and their needs for addressing harm. My research highlights a survivor's perspective of understanding harm that centers on needs (Chapter 4, 5). This framing helps survivors in their sensemaking of harm and needs, and helps to design harm reparations according to survivors' needs. For example, survivors consider their relationship with perpetrators because it influences their decisions about disclosure and forgiveness. Survivors are often concerned with whether harm occurred in public or private spaces, as this impacts the harm's reach and the impact. Survivors may consider multiple encounters of harm together as an ongoing experience. These findings underscore the importance of framing and describing harm from users' perspectives and explore this need-based understanding systematically to better capture what matters to those affected.

7.2 Adapting Offline Justice Models for Online Contexts

In this section, I use my research as an example to discuss key considerations for adapting justice frameworks for online spaces. This process is not straightforward: it requires careful evaluation of incentives, resources and community's notion of justice.

Incentive structures

Offline restorative justice frameworks typically engage survivors and perpetrators as primary stakeholders, each motivated to participate by distinct needs. Survivors may seek truth-telling, empowerment, or restitution, while perpetrators may view it as an alternative to punitive measures or a means of repairing harm [184]. These conversations are often mediated

by professional restorative justice facilitators. Recognizing that incentive structures differ in online contexts, my research first identified the motivations of online stakeholders before designing solutions.

For online harm survivors, I identified a variety of incentives that align with restorative justice practices, including the need to make sense of the harm, seek emotional support, ensure personal safety, hold perpetrators accountable, and foster a more positive online environment (Chapter 4). However, my findings reveal that there are limited or no incentives for perpetrators to participate in online restorative justice processes due to the absence of accountability structure in online environments (Chapter 6). Some moderators, who approach harm with a restorative mindset, also face barriers: their current practices are predominantly punitive, constrained by a lack of resources, tools, and awareness of restorative alternatives (Chapter 6).

Given these differing incentives, my research explores restorative justice across distinct stages of harm resolution for survivors, perpetrators, and moderators. For survivors, who are central to restorative justice and most prepared to embrace new practices, my work prioritizes designing solutions tailored to their needs and motivations, where I created SnuggleSense to provide survivors a structured sensemaking process in addressing harm (Chapter 5). Justice practices can begin with stakeholders who have the strongest incentives and where resources are most accessible, gradually transitioning to more labor- and resource-intensive practices as broader willingness and accountability structures develop (Chapter 6).

For perpetrators, establishing incentives requires creating robust online accountability mechanisms. These mechanisms could foster a sense of responsibility and increase the likelihood of participation in restorative practices (Chapter 6). For moderators, who often express motivation for restorative justice but are limited by resources, my research suggests starting with existing moderation practices. Embedding restorative justice principles, language, and framing into current workflows can provide a feasible pathway to shift from punitive to restorative approaches over time (Chapter 6).

Resource availability

Implementing new justice-inspired practices online necessitates careful consideration of the resources available for transition and implementation. Offline restorative justice practices are often labor-intensive, requiring the involvement of diverse parties and trained facilitators to mediate conversations [184]. Given the vast scale of online harm and the limited moderation resources available, it can be challenging to directly implement this practice online.

My research designs support mechanisms for survivors that account for resource constraints, while also identifying what resources are essential for effective restorative justice practices. For example, SnuggleSense provides survivors with a structured sensemaking process with guidance and informational support from other survivors (Chapter 5). By guiding survivors toward available resources and fostering supportive communities, my work demonstrates how survivors can achieve empowerment even in the absence of direct human intervention.

My research also presents ways for online communities to start with partial restorative justice practices that stem from existing moderation practices such as changing moderation explanations to focus on impact instead of rules (Chapter 6). In addition, I discussed resources needed to develop more comprehensive restorative justice practices such as establishing perpetrator accountability and expanding peoples' awareness of alternative justice models (Chapter 6). My research outlines a path for incrementally evolving online restorative justice practices with consideration of available resources.

Community notion of justice

Adapting justice models to online contexts requires an understanding of existing community perspectives on justice. My research shows that survivors, perpetrators, and moderators often default to a punitive view of justice, perceiving punishment as the primary—and sometimes the only—solution (Chapter 6). Survivors, for example, may believe that content moderation is the only option they can resort to; some perpetrators may accept punishment as deserved; moderators frequently see punitive measures as necessary for reform.

Shifting these perceptions can be approached in two ways. First, it is through engaging people in the process of imagining and designing harm resolutions. Rather than enforcing specific actions from various stakeholders, my research utilizes sensemaking to broaden perspectives on harm resolution options for survivors inspired by restorative justice (Chapter 5). Over time, SnuggleSense could act as an educational tool, enabling survivors, moderators, perpetrators, policymakers, and researchers to explore and adopt alternative approaches to resolving harm.

Second, small-scale implementations offer a pathway to change. In my research, some moderators and community members with prior exposure to restorative justice have already begun integrating these ideas into their practices (Chapter 6). These early trials provide tangible examples of how restorative justice can function effectively, fostering interest and trust among stakeholders and laying the groundwork for broader adoption in online spaces.

7.3 Concluding Thoughts: Toward Diverse Justice Approaches in Digital Space

Societies have long debated justice and appropriate responses to harm before the internet transformed how we interact. The internet, however, introduces new challenges, with harm occurring at an unprecedented scale, speed, and in novel forms. While early online communities experimented with various approaches to addressing harm, the rise of large platforms has led to a default reliance on punitive measures.

I argue, however, that we must continually reimagine what justice can look like in evolving digital spaces. Restorative justice is one of many justice frameworks flourishing in society today—but it is not the final destination. Our goal should be to ensure that punishment is not the sole or default response to online harm. Instead, we must embrace a diversity of

approaches that account for the specific nature of the harm and the individualized needs of those affected. This approach fosters a more nuanced and evolving understanding of justice, one that keeps pace with society's growth and the intricacies of the online world.

Bibliography

[1] Achini Adikari et al. "A cognitive model for emotion awareness in industrial Chatbots". In: *2019 IEEE 17th international conference on industrial informatics (INDIN)*. Vol. 1. IEEE. 2019, pp. 183–186.

[2] Sonam Adinolf and Selen Turkay. "Toxic behaviors in Esports games: player perceptions and coping strategies". In: *Proceedings of the 2018 Annual Symposium on computer-human interaction in play companion extended abstracts*. 2018, pp. 365–372.

[3] Ivo Aertsen, Daniela Bolívar, Nathalie Lauwers, et al. "Restorative justice and the active victim: exploring the concept of empowerment". In: *Temida* 14.1 (2011), pp. 5–19.

[4] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–236. DOI: 10.1257/jep.31.2.211. URL: http://pubs.aeaweb.org/doi/10.1257/jep.31.2.211.

[5] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. "Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–30.

[6] Nazanin Andalibi and Patricia Garcia. "Sensemaking and coping after pregnancy loss: the seeking and disruption of emotional validation online". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–32.

[7] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. "Sensitive self-disclosures, responses, and social support on Instagram: The case of# depression". In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017, pp. 1485–1500.

[8] Monica Anderson. "A majority of teens have experienced some form of cyberbullying". In: (2018).

[9] Zahra Ashktorab and Jessica Vitak. "Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 3895–3905.

[10] Geoffrey C Barnes et al. "Are restorative justice conferences more fair than criminal courts? Comparing levels of observed procedural justice in the reintegrative shaming experiments (RISE)". In: *Criminal Justice Policy Review* 26.2 (2015), pp. 103–130.

[11] Kristen Barta, Katelyn Wolberg, and Nazanin Andalibi. "Similar Others, Social Comparison, and Social Support in Online Support Groups". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023), pp. 1–35.

[12] Gordon Bazemore and Mara Schiff. *Restorative community justice: Repairing harm and transforming communities*. Routledge, 2015.

[13] S Gordon Bazemore. *A comparison of four restorative conferencing models*. US Department of Justice, Office of Justice Programs, Office of Juvenile ..., 2001.

[14] S Gordon Bazemore and Lode Walgrave. *Restorative juvenile justice: Repairing the harm of youth crime*. Criminal Justice Press Monsey, NY, 1999.

[15] Derek Bell. *And we are not saved: The elusive quest for racial justice*. Basic Books, 2008.

[16] Nicole A Beres et al. "Don't you know that you're toxic: Normalization of toxicity in online gaming". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–15.

[17] Patrick Biernacki and Dan Waldorf. "Snowball Sampling: Problems and Techniques of Chain Referral Sampling". In: *Sociological Methods & Research* 10.2 (Nov. 1981). Publisher: SAGE PublicationsSage CA: Los Angeles, CA, pp. 141–163. ISSN: 0049-1241. DOI: 10.1177/004912418101000205. URL: http://journals.sagepub.com/doi/10.1177/004912418101000205.

[18] Reuben Binns et al. "Like trainer, like bot? Inheritance of bias in algorithmic content moderation". In: *International Conference on Social Informatics*. Springer. 2017, pp. 405–415.

[19] Lindsay Blackwell et al. "Classification and Its Consequences for Online Harassment: Design Insights from HeartMob." In: *PACMHCI* 1.CSCW (2017), pp. 24–1.

[20] Lindsay Blackwell et al. "When Online Harassment is Perceived as Justified". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[21] Blizzard. *Blizzard's In-Game Code of Conduct*. en-us. 2020. URL: https://us.battle.net/support/en/article/42673 (visited on 10/04/2020).

[22] Jane Bolitho and Jasmine Bruce. "Science, art and alchemy: Best practice in facilitating restorative justice". In: *Contemporary Justice Review* 20.3 (2017), pp. 336–362.

[23] Shannon A Bowen. "Using classic social media cases to distill ethical guidelines for digital engagement". In: *Journal of Mass Media Ethics* 28.2 (2013), pp. 119–133.

[24] Lynn S Branham. *Eradicating the Label "Offender" from the Lexicon of Restorative Practices and Criminal Justice*. Aug. 2019. URL: http://wakeforestlawreview.com/2019/08/eradicating-the-label-offender-from-the-lexicon-of-restorative-practices-and-criminal-justice/.

[25] Jie Cai and Donghee Yvette Wohn. "What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives". In: *Conference companion publication of the 2019 on computer supported cooperative work and social computing*. 2019, pp. 166–170.

[26] Caitlin Ring Carlson and Luc S Cousineau. "Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice". In: *Journal of Media Ethics* 35.4 (2020), pp. 202–213.

[27] Archie B Carroll. "Carroll's pyramid of CSR: taking another look". In: *International journal of corporate social responsibility* 1.1 (2016), pp. 1–8.

[28] Eshwar Chandrasekharan et al. "Crossmod: A cross-community learning-based system to assist reddit moderators". In: *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019), pp. 1–30.

[29] Eshwar Chandrasekharan et al. "The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), p. 32.

[30] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage, 2006.

[31] Gary Charness, Uri Gneezy, and Michael A Kuhn. "Experimental methods: Between-subject and within-subject design". In: *Journal of economic behavior & organization* 81.1 (2012), pp. 1–8.

[32] David M Chavis and Abraham Wandersman. "Sense of community in the urban environment: A catalyst for participation and community development". In: *American journal of community psychology* 18.1 (1990), pp. 55–81.

[33] Janet X Chen et al. "Trauma-informed computing: Towards safer technology experiences for all". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–20.

[34] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial behavior in online discussion communities". In: *Ninth International AAAI Conference on Web and Social Media*. 2015.

[35] Alexander Cho et al. "The" Comadre" Project: An Asset-Based Design Approach to Connecting Low-Income Latinx Families to Out-of-School Learning Opportunities". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–14.

[36] Danielle Keats Citron. "Addressing Cyber Harassment: An Overview of Hate Crimes in Cyberspace". en. In: Rochester, NY: Social Science Research Network, 2015. URL: https://papers.ssrn.com/abstract=2932358 (visited on 05/07/2020).

[37] Danielle Keats Citron and Mary Anne Franks. "Criminalizing Revenge Porn". In: *Wake Forest Law Review* 49 (2014). URL: http://heinonline.org/HOL/Page?handle=hein.journals/wflr49%7B%5C&%7Did=357%7B%5C&%7Ddiv=15%7B%5C&%7Dcollection=journals.

[38] Erin Daly. "Transformative justice: Charting a path to reconciliation". In: *Int'l Legal Persp.* 12 (2001), p. 73.

[39] Kathleen Daly. "Conferencing in Australia and New Zealand: Variations, research findings and prospects". In: *Restorative justice for juveniles: Conferencing, mediation and circles* (2001), pp. 59–89.

[40] Kathleen Daly. "Restorative justice: The real story". en. In: *Punishment & Society* 4.1 (Jan. 2002), pp. 55–79. ISSN: 1462-4745, 1741-3095. DOI: 10.1177/14624740222228464. URL: http://journals.sagepub.com/doi/10.1177/14624740222228464 (visited on 02/14/2023).

[41] Angela Dean and Daniel Voss. *Design and analysis of experiments*. Springer, 1999.

[42] Lina Dencik, Arne Hintz, and Jonathan Cable. "Towards data justice". In: *DATA POLITICS* (2017), p. 167.

[43] James Dignan. *Understanding victims and restorative justice*. McGraw-Hill Education (UK), 2004.

[44] Jill P Dimond et al. "Hollaback! The role of storytelling online in a social movement organization". In: *Proceedings of the 2013 conference on Computer supported cooperative work*. 2013, pp. 477–490.

[45] Judith S Donath. "Identity and deception in the virtual community". In: *Communities in cyberspace*. Routledge, 2002, pp. 37–68.

[46] Bryan Dosono and Bryan Semaan. "Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–13.

[47] Maeve Duggan. *Online harassment*. Pew Research Center, 2014.

[48] Maeve Duggan. *Online Harassment 2017*. en-US. July 2017. URL: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/ (visited on 12/14/2019).

[49] Maeve Duggan and Aaron Smith. "6% of online adults are reddit users". In: (2013).

[50] Lee B Erickson et al. "The boundaries between: Parental involvement in a teen's online world". In: *Journal of the Association for Information Science and Technology* 67.6 (2016), pp. 1384–1403.

[51] Matthew Evans. "Structural Violence, Socioeconomic Rights, and Transformative Justice". In: *Journal of Human Rights* 15.1 (Jan. 2016), pp. 1–20. ISSN: 1475-4835. DOI: 10.1080/14754835.2015.1032223.

[52] Stephen B Fawcett et al. "A contextual-behavioral model of empowerment: Case studies involving people with physical disabilities". In: *American Journal of Community Psychology* 22.4 (1994), pp. 471–496.

[53] Casey Fiesler and Brianna Dym. "Moving Across Lands: Online Platform Migration in Fandom Communities". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW1 (May 2020). DOI: 10.1145/3392847. URL: https://doi.org/10.1145/3392847.

[54] Michel Foucault. *Discipline and punish: The birth of the prison*. Vintage, 2012. ISBN: 0-307-81929-9.

[55] Nancy Fraser. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy". In: *Social Text* 25/26 (1990), pp. 56–80. ISSN: 0164-2472. DOI: 10.2307/466240. JSTOR: 466240.

[56] Diana Freed et al. "Understanding Digital-Safety Experiences of Youth in the U.S." In: (2023).

[57] David Garland. *Punishment and modern society: A study in social theory*. University of Chicago Press, 1993.

[58] Theo Gavrielides. "Collapsing the labels 'victim'and 'offender'in the Victims' Directive and the paradox of restorative justice". In: *Restorative Justice* 5.3 (2017), pp. 368–381.

[59] Theo Gavrielides. *Restorative justice: Ideals and realities*. Routledge, 2017.

[60] Alix Gerber. "Participatory speculation: Futures of public safety". In: *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2*. 2018, pp. 1–4.

[61] Tarleton Gillespie. "Content moderation, AI, and the question of scale". In: *Big Data & Society* 7.2 (2020), p. 2053951720943234.

[62] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[63] Tarleton Gillespie. "Governance of and by platforms". In: *Sage handbook of social media* (2017).

[64] Robert Gorwa. "What is platform governance?" In: *Information, Communication & Society* (2019), pp. 1–18.

[65] Nitesh Goyal, Leslie Park, and Lucy Vasserman. "" You have to prove the threat is real": Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–17.

[66] Why Is This Happening? *Thinking about how to abolish prisons with Mariame Kaba: podcast; transcript*. Apr. 2019. URL: https://www.nbcnews.com/think/opinion/thinking-about-how-abolish-prisons-mariame-kaba-podcast-transcript-ncna992721.

[67] A Hasinoff, AD Gibson, and N Salehi. "The promise of restorative justice in addressing online harm". In: *Tech Stream* 27 (2020).

[68] Amy A Hasinoff and Nathan Schneider. "From Scalability to Subsidiarity in Addressing Online Harm". In: *Social Media+ Society* 8.3 (2022), p. 20563051221126041.

[69] Caroline Haythornthwaite. "Strong, weak, and latent ties and the impact of new media". In: *The information society* 18.5 (2002), pp. 385–401.

[70] Natali Helberger, Jo Pierson, and Thomas Poell. "Governing online platforms: From contested to cooperative responsibility". In: *The information society* 34.1 (2018), pp. 1–14.

[71] Zorah Hilvert-Bruce and James T Neill. "I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming". In: *Computers in Human Behavior* 102 (2020), pp. 303–311.

[72] Belinda Hopkins. "Restorative justice in schools". In: *Support for Learning* 17.3 (2002), pp. 144–149.

[73] Youyang Hou et al. "Factors in fairness and emotion in online case resolution systems". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 2511–2522.

[74] Maggie Hughes and Deb Roy. "Keeper: An Online Synchronous Conversation Environment Informed by In-Person Facilitation Practices". In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 2020, pp. 275–279.

[75] Jon Hurwitz and Mark Peffley. "Explaining the great racial divide: Perceptions of fairness in the US criminal justice system". In: *The Journal of Politics* 67.3 (2005), pp. 762–783.

[76] Jane Im et al. "Women's Perspectives on Harm and Justice after Online Harassment". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–23.

[77] Jane Im et al. "Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–18.

[78] Joshua E Introne. *Adaptive mediation in groupware*. Brandeis University, 2008.

[79] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. "Does transparency in moderation really matter? User behavior after content removal explanations on reddit". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–27.

[80] Shagun Jhaver, Larry Chan, and Amy Bruckman. "The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action". In: *First Monday* 23.2 (2018). URL: http://firstmonday.org/ojs/index.php/fm/article/view/8232.

[81] Shagun Jhaver et al. ""Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359294. URL: https://doi.org/10.1145/3359294.

[82] Shagun Jhaver et al. "Online Harassment and Content Moderation: The Case of Blocklists". In: *ACM Trans. Comput.-Hum. Interact.* 25.2 (Mar. 2018). ISSN: 1073-0516. DOI: 10.1145/3185593. URL: https://doi.org/10.1145/3185593.

[83] Jialun Aaron Jiang et al. "Moderation challenges in voice-based online communities on discord". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–23.

[84] Brett Gregory Johnson. "Speech, harm, and the duties of digital intermediaries: Conceptualizing platform ethics". In: *Journal of Media Ethics* 32.1 (2017), pp. 16–27.

[85] Gerry Johnstone and Daniel Van Ness. *Handbook of Restorative Justice*. en. Google-Books-ID: U2UQBAAAQBAJ. Routledge, Jan. 2013. ISBN: 978-1-134-01519-1.

[86] Gerry Johnstone and Daniel Van Ness. *Handbook of restorative justice*. Routledge, 2013.

[87] Mariame Kaba. *We Do This Til We Free Us: Abolitionist Organizing and Transforming Justice*. Haymarket Books, 2021.

[88] Mariame Kaba and Naomi Murakawa. *We Do this' til We Free Us: Abolitionist Organizing and Transforming Justice*. Haymarket Books, 2021.

[89] David R. Karp and Beau Breslin. "Restorative Justice in School Communities". en. In: *Youth & Society* 33.2 (Dec. 2001), pp. 249–272. ISSN: 0044-118X. DOI: 10.1177/0044118X01033002006. URL: https://doi.org/10.1177/0044118X01033002006 (visited on 12/10/2019).

[90] Naveena Karusala et al. "Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-making in Public Services". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–16.

[91] Charles Kiene and Benjamin Mako Hill. "Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–8.

[92] Sara Kiesler, Robert Kraut, and Paul Resnick. "Regulating behavior in online communities". In: *Building Successful Online Communities: Evidence-Based Social Design* (2012).

[93] Mimi E. Kim. "From carceral feminism to transformative justice: Women-of-color feminism and alternatives to incarceration". In: *Journal of Ethnic & Cultural Diversity in Social Work* 27.3 (2018), pp. 219–233. DOI: 10.1080/15313204.2018.1474827. eprint: https://doi.org/10.1080/15313204.2018.1474827. URL: https://doi.org/10.1080/15313204.2018.1474827.

[94] Anna King and Shadd Maruna. "Is a conservative just a liberal who has been mugged? Exploring the origins of punitive views". In: *Punishment & Society* 11.2 (2009), pp. 147–169.

[95] Yubo Kou. "Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–21.

[96] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[97] Travis Kriplean et al. "Supporting reflective public thought with considerit". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. Seattle, Washington, USA: ACM Press, 2012, pp. 265–274. URL: http://dl.acm.org/citation.cfm?doid=2145204.2145249.

[98] Etienne G Krug et al. "The world report on violence and health". In: *The lancet* 360.9339 (2002), pp. 1083–1088.

[99] Deepak Kumar et al. *Understanding Longitudinal Behaviors of Toxic Accounts on Reddit*. en. arXiv:2209.02533 [cs]. Sept. 2022. URL: http://arxiv.org/abs/2209.02533 (visited on 02/27/2023).

[100] Shih-Ya Kuo, Dennis Longmire, and Steven J Cuvelier. "An empirical assessment of the process of restorative justice". In: *Journal of Criminal Justice* 38.3 (2010), pp. 318–328.

[101] Noam Lapidot-Lefler and Azy Barak. "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition". In: *Computers in human behavior* 28.2 (2012), pp. 434–443.

[102] Jeff Latimer, Craig Dowden, and Danielle Muise. "The Effectiveness of Restorative Justice Practices: A Meta-Analysis". In: *The Prison Journal* 85.2 (June 2005). Publisher: SAGE Publications Inc, pp. 127–144. ISSN: 0032-8855. DOI: 10.1177/0032885505276969. URL: https://doi.org/10.1177/0032885505276969 (visited on 08/29/2020).

[103] Glenn Laverack. "Improving health outcomes through community empowerment: a review of the literature". In: *Journal of Health, Population and Nutrition* (2006), pp. 113–120.

[104] Jooyoung Lee et al. "Online Self-Disclosure, Social Support, and User Engagement During the COVID-19 Pandemic". In: *Trans. Soc. Comput.* 6.3–4 (Dec. 2023). DOI: 10.1145/3617654. URL: https://doi.org/10.1145/3617654.

[105] Amanda Lenhart et al. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute, 2016.

[106] Hanlin Li, Lynn Dombrowski, and Erin Brady. "Working toward empowering a community: How immigrant-focused nonprofit organizations use Twitter during political conflicts". In: *Proceedings of the 2018 ACM International Conference on Supporting Group Work*. 2018, pp. 335–346.

[107] Wookjae Maeng and Joonhwan Lee. "Designing and Evaluating a Chatbot for Survivors of Image-Based Sexual Abuse". In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–21. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3517629.

[108] Kaitlin Mahar, Amy X. Zhang, and David Karger. "Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174160.

[109] Alice E Marwick. "Morally motivated networked harassment as normative reinforcement". In: *Social Media+ Society* 7.2 (2021), p. 20563051211021378.

[110] Abraham Harold Maslow. "A theory of human motivation." In: *Psychological review* 50.4 (1943), p. 370.

[111] Alison Mathie and Gord Cunningham. "From clients to citizens: Asset-based community development as a strategy for community-driven development". In: *Development in practice* 13.5 (2003), pp. 474–486.

[112] Alison Mathie and Gord Cunningham. "Who is driving development? Reflections on the transformative potential of asset-based community development". In: *Canadian Journal of Development Studies/Revue canadienne d'études du développement* 26.1 (2005), pp. 175–186.

[113] David C McClelland. *Human motivation*. CUP Archive, 1987.

[114] Paul McCold. "Toward a holistic vision of restorative juvenile justice: A reply to the maximalist model". In: *Contemporary Justice Review* 3.4 (2000), pp. 357–414.

[115] Sharan B Merriam and Robin S Grenier. *Qualitative research in practice: Examples for discussion and analysis*. Jossey-Bass, 2019.

[116] Jon'a F Meyer. "History repeats itself: Restorative justice in Native American communities". In: *Journal of Contemporary Criminal Justice* 14.1 (1998), pp. 42–57.

[117] Daniel C Molden. "Understanding priming effects in social psychology: What is "social priming" and how does it occur?" In: *Social cognition* 32.Supplement (2014), pp. 1–11.

[118] Sassy Molyneux et al. "Community accountability at peripheral health facilities: a review of the empirical literature and development of a conceptual framework". In: *Health policy and planning* 27.7 (2012), pp. 541–554.

[119] Allison Morris, Gabrielle M Maxwell, and Jeremy P Robertson. "Giving victims a voice: A New Zealand experiment". In: *The Howard Journal of Criminal Justice* 32.4 (1993), pp. 304–321.

[120] Garrett Morrow et al. "The emerging science of content labeling: Contextualizing social media content moderation". In: *Journal of the Association for Information Science and Technology* 73.10 (2022), pp. 1365–1386.

[121] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. "Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–17.

[122] Sarah Myers West. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms". en. In: *New Media & Society* 20.11 (Nov. 2018), pp. 4366–4383. ISSN: 1461-4448, 1461-7315. DOI: 10.1177/1461444818773059. URL: http://journals.sagepub.com/doi/10.1177/1461444818773059 (visited on 07/17/2020).

[123] Angela Nagle. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing, 2017.

[124] Safiya Umoja Noble. "Algorithms of oppression: How search engines reinforce racism". In: *Algorithms of oppression*. New York university press, 2018.

[125] Elinor Ostrom. *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge, 1990.

[126] Richard Stanley Peters. *The concept of motivation*. Routledge, 2015.

[127] Whitney Phillips. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press, 2015.

[128] Kay Pranis. *Little book of circle processes: A new/old approach to peacemaking*. Simon and Schuster, 2015.

[129] J Rappaport. "Terms of Empowerment: Theories for Community Psychology". In: *American Journal of Community Psychology* 15.2 (1987), pp. 122–144.

[130] Elizabeth Reid. "Hierarchy and power". In: *Communities in cyberspace* (1999), pp. 107–133.

[131] Sarah T Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.

[132] Sarah T Roberts. "Digital detritus:'Error'and the logic of opacity in social media content moderation". In: *First Monday* 23.3 (2018).

[133] Oliver C Robinson. "Sampling in interview-based qualitative research: A theoretical and practical guide". In: *Qualitative research in psychology* 11.1 (2014), pp. 25–41.

[134] Aja Romano. *How the alt-right uses internet trolling to confuse you into dismissing its ideology*. 2017. URL: https://www.vox.com/2016/11/23/13659634/alt-right-trolling (visited on 03/26/2018).

[135] Jon Ronson. "How one stupid tweet blew up Justine Sacco's life". In: *The New York Times Magazine* 12 (2015).

[136] Johnny Saldaña. *The coding manual for qualitative researchers*. sage, 2021.

[137] Niloufar Salehi. "Do no harm". In: *Logic Magazine* (2020).

[138] Niloufar Salehi et al. "Communicating context to the crowd for complex writing tasks". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 1890–1901.

[139] santaclaraprinciples.org. *Santa Clara principles on transparency and accountability in content moderation*. 2022. URL: https://santaclaraprinciples.org/.

[140] Susan M Sawyer et al. "The age of adolescence". In: *The Lancet Child & Adolescent Health* 2.3 (2018), pp. 223–228.

[141] Devansh Saxena and Shion Guha. "Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making". In: *ACM Journal on Responsible Computing* 1.1 (2024), pp. 1–32.

[142] Morgan Klaus Scheuerman et al. "A Framework of Severity for Harmful Content Online". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–33.

[143] Hanna Schneider et al. "Empowerment in HCI-A survey and framework". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–14.

[144] Sarita Schoenebeck and Lindsay Blackwell. "Reimagining social media governance: Harm, accountability, and repair". In: *Yale JL & Tech.* 23 (2020), p. 113.

[145] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. "Drawing from justice theories to support targets of online harassment". In: *new media & society* 23.5 (2021), pp. 1278–1300.

[146] Sarita Schoenebeck et al. *Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries*. Jan. 2023. DOI: 10.1145/3544548.3581020. arXiv: 2301.11715 [cs].

[147] Carol F Scott et al. "Trauma-informed social media: Towards solutions for reducing and healing online harm". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–20.

[148] Joseph Seering. "Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–28.

[149] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. "Metaphors in moderation". In: *New Media & Society* (2020), p. 1461444820964968.

[150] Joseph Seering, Robert Kraut, and Laura Dabbish. "Shaping pro and anti-social behavior on twitch through moderation and example-setting". In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017, pp. 111–125.

[151] Joseph Seering et al. "Moderator engagement and community development in the age of algorithms". In: *New Media & Society* (2019), p. 1461444818821316.

[152] Renee Shelby et al. "Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 723–741.

[153] Donna Spencer. *Card sorting: Designing usable categories*. Rosenfeld Media, 2009.

[154] Timo Spring et al. "Empathic Response Generation in Chatbots." In: *SwissText*. 2019.

[155] Miriah Steiger et al. "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380966. URL: https://doi.org/10.1145/3411764.3445092.

[156] Heather Strang and Lawrence W Sherman. "Repairing the harm: Victims and restorative justice". In: *Utah L. Rev.* (2003), p. 15.

[157] Sharifa Sultana et al. "'ShishuShurokkha': A Transformative Justice Approach for Combating Child Sexual Abuse in Bangladesh". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–23.

[158] Sharifa Sultana et al. "Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger". In: Mar. 2021. DOI: 10.1145/3411764.3445154.

[159] Nicolas Suzor. "Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms". In: *Social Media + Society* 4.3 (2018), p. 2056305118787812.

[160] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. "Evaluating the legitimacy of platform governance: A review of research and a shared research agenda". In: *International Communication Gazette* 80.4 (2018), pp. 385–400.

[161] Nicolas P Suzor et al. "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation". In: *International Journal of Communication* 13 (2019), p. 18.

[162] Zenon Szablowinski. "Punitive justice and restorative justice as social reconciliation". In: *The Heythrop Journal* 49.3 (2008), pp. 405–422.

[163] Samuel Hardman Taylor et al. "Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–26.

[164] Kurt Thomas et al. ""It's Common and a Part of Being a Content Creator": Understanding How Creators Experience and Cope with Hate and Harassment Online". In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–15. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3501879.

[165] Alexandra To et al. "" They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–29.

[166] Daniel Van Ness and Karen Heetderks Strong. *Restoring justice: An introduction to restorative justice*. Routledge, 2014.

[167] Daniel W Van Ness. "An overview of restorative justice around the world". In: (2016).

[168] Jonathan Vanian. *Twitter Toughens Rules on Nudity and Revenge Porn — Fortune*. 2017. URL: http://fortune.com/2017/10/27/nudity-revenge-porn-twitter/ (visited on 03/26/2018).

[169] Jessica Vitak et al. "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*. 2017, pp. 1231–1245.

[170] Emily Vogels. "The state of online harassment". In: *Pew Research Center* (2021).

[171] Ron Wakkary et al. "Material speculation: Actual artifacts for critical inquiry". In: *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*. 2015, pp. 97–108.

[172] Noel Warford et al. *SoK: A Framework for Unifying At-Risk User Research*. Dec. 2021. arXiv: arXiv:2112.07047.

[173] Charlie Warzel. "A Honeypot for Assholes: Inside Twitters 10-Year Failure to Stop Harassment". In: *BuzzFeed News* (2016).

[174] Karl E Weick. *Sensemaking in organizations*. Vol. 3. Sage, 1995.

[175] Michael Wenzel et al. "Retributive and restorative justice". In: *Law and human behavior* 32.5 (2008), pp. 375–389.

[176] Sarah Myers West. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms". In: *New Media & Society* (2018).

[177] Pamela J Wisniewski et al. "Adolescent online safety: the" moral" of the story". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 1258–1271.

[178] Donghee Yvette Wohn. "Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–13.

[179] Richmond Y Wong and Tonya Nguyen. "Timelines: A World-Building Activity for Values Advocacy". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.

[180] William R Wood and Masahiro Suzuki. "Four challenges in the future of restorative justice". In: *Victims & Offenders* 11.1 (2016), pp. 149–172.

[181] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. "Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–15.

[182] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. "Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach". In: *ACM Transactions on Computer-Human Interaction* (2023).

[183] Mireia Yurrita et al. "Towards a multi-stakeholder value-based assessment framework for algorithmic systems". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 535–563.

[184] Howard Zehr. *The little book of restorative justice: Revised and updated*. Simon and Schuster, 2015.

[185] Yi Zeng et al. "Ai risk categorization decoded (air 2024): From government regulations to corporate policies". In: *arXiv preprint arXiv:2406.17864* (2024).

[186] Marc A Zimmerman. "Empowerment theory". In: *Handbook of community psychology*. Springer, 2000, pp. 43–63.