Addressing harm in online gaming communities – the opportunities and challenges for a restorative justice approach

SIJIA XIAO, University of California, Berkeley, USA
SHAGUN JHAVER, Rutgers University, USA
NILOUFAR SALEHI, University of California, Berkeley, USA

Most platforms implement some form of content moderation to address interpersonal harms such as harassment. Content moderation relies on offender-centered, punitive justice approaches such as bans and content removals. We consider an alternative justice framework, restorative justice, which aids victims to heal, supports offenders to repair the harm, and engages community members to address the harm collectively. To understand the utility of restorative justice in addressing online harm, we interviewed 23 users from Overwatch gaming communities, including moderators, victims, and offenders. We understand how they currently handle harm cases through the lens of restorative justice and identify their attitudes toward implementing restorative justice processes. Our analysis reveals that while online communities have needs for and existing structures to support restorative justice, there are structural, cultural, and resource-related obstacles to implementing this new approach within the existing punitive framework. We discuss the opportunities and challenges for applying restorative justice in online spaces.

CCS Concepts: • Human-centered computing → Empirical studies in collaborative and social computing.

Additional Key Words and Phrases: content moderation, online harassment, alternative justice, Discord, Overwatch

ACM Reference Format:
Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2022. Addressing harm in online gaming communities – the opportunities and challenges for a restorative justice approach. J. ACM 37, 4, Article 111 (November 2022), 35 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Social media platforms most frequently address online harms such as harassment through content moderation, which involves the review and removal of user-submitted content that violate the platform’s rules [39, 83]. However, despite efforts in research and industry to develop new moderation practices in recent years, the number of people experiencing severe forms of harassment has grown. In 2014, 15% of Americans reported experiencing severe harassment, including physical threats, stalking, sexual harassment, and sustained harassment [31]. That number grew to 18% in 2017, and 25% in 2021. Meanwhile, a growing number of people have reported simultaneously experiencing multiple forms of severe harassment [32, 110]. Research shows that online harms
are insufficiently addressed by platforms’ and communities’ current approaches [21, 33, 70, 111] and 79% of Americans say that social media companies are doing a fair or poor job at addressing online harassment [110]. Alternative approaches are desperately needed, but what should the guiding goals of such approaches be and what might they look like in practice?

Restorative justice is a philosophy and practice of justice that argues for working to repair harm and restore communities after harm has occurred. In this paper, we apply a restorative justice lens to study how an online gaming community currently addresses—and might alternatively address—severe online harm. We chose to focus on restorative justice here because it has an established offline practice and has been successfully institutionalized to address harm in other contexts [4, 66, 68]. Restorative justice focuses on providing care, support, and meeting peoples’ needs after harm has occurred. It has three major principles: (1) 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 [72, 117]. In practice, restorative justice addresses harm differently than more common punitive models. The main tool for action in a punitive justice model, as in content moderation, is punishing the person who has violated the rules. In contrast, the main tool for action in restorative justice is communication involving the harmed person, the offender, and the community. For instance, in a common restorative justice practice called a victim-offender conference, the victim and the offender meet to discuss how to address the harm under the guidance of a facilitator. Interested community members are also invited to join this conversation. The conference aims to address the needs and obligations of all parties involved, including the victim, the offender and the community members. The follow-up process of the conference may include apologies or community service by the offenders [117].

In this paper, we use 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. This research has two major goals: first, we explore how the victims, offenders, and moderators currently experience or address online harm. We use the three restorative justice principles to evaluate current practices and identify the potential needs for restorative justice practices. Second, we introduce the victim-offender conference to participants to learn about their perspectives of one key restorative justice process. We use this as a way to understand the promises and challenges of practicably implementing restorative justice practices in an online setting.

We study harm cases in the Overwatch gaming community in two major platforms that the members of this community use: Overwatch 1 and Discord 2. This is a community that frequently issues moderation sanctions and is considered to have a “toxic” environment by many players [20, 106]. 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 pre-conferencing, a common practice in restorative justice to learn people’s history and preferences, explain restorative justice to them, and evaluate the appropriateness of holding a victim-offender conference [117]. Additionally, given that restorative justice has mostly been developed through practice[107], we deepened our understanding of its principles and practices by conducting two interviews with restorative justice practitioners.

We found that current, punitive online moderation processes do not effectively stop the perpetuation of harm. First, content moderation is offender-centered and does not address victims’ needs such as receiving support or healing from the harm. Though victims may report individual

1https://playoverwatch.com
2https://discord.com

J. ACM, Vol. 37, No. 4, Article 111. Publication date: November 2022.
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 harm they cause. When punishment is not effective, 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 by reacting punitively toward perpetrators or victims.

We found that some moderation practices align with restorative justice, and a few participants have attempted implementing restorative justice practices in their own 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 restorative justice practices in online communities.

We found that some moderation practices align with restorative justice, and a few participants have attempted implementing restorative justice practices in their own 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 restorative justice practices in online communities.

Our work contributes to a growing line of research that applies alternative justice frameworks to address online harm [21, 26, 43, 44, 77, 90, 93]. By examining how we can adapt successful offline restorative justice practices to online communities, our work sheds light on ways to address the needs of key stakeholders that go beyond maintaining healthy community content and working within a perpetrator-oriented model. We highlight how restorative justice has the potential to reduce the perpetuation of harm and improve community culture in the long run.

2 RELATED WORK

In this section, we review related work on online harm and content moderation. We also describe the restorative justice framework and its current applications.

2.1 Online Harm and Content Moderation

Online harm can refer to a myriad of toxic behaviors, such as public shaming [87], trolling [19], and invasion of privacy [70]. In this paper, most of the harm we study falls under Duggan’s definition of online harassment, which includes six categories: offensive name-calling, purposeful embarrassment, physical threats, sustained harassment, stalking, and sexual harassment [32]. Online gaming communities are particularly susceptible to online harm: nearly three-quarters of all online game players have suffered some form of online abuse [69]. Researchers have found that harm in games may intersect with harm towards certain groups, for example, females [98], gender minorities [88] or people of color [41].

Online platforms currently address harm through content moderation, which usually involves punitive measures such as removal of content, muting, or bans [39, 83]. Online platforms’ moderators can be volunteers who are users of the platform or commercial content moderators hired by social media companies [83, 97]. In recent years, social media platforms have also begun using automated, AI-based tools such as bots to help enact moderation [8, 59]. Current online moderation widely applies the graduated sanction model: the moderation of offending behaviors begins with persuasion and proceeds to more forceful measures [60, 80]. 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. Kiesler and Kraut argue that graduated sanctions are an effective model to regulate offenders’ behaviors [60].

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 communities. 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 [30, 83, 115]. 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) [2], alt-right trolls [79, 86] and revenge porn [22, 109]. Researchers have examined ways to improve the current moderation processes, for example, through setting positive examples and social norms [17, 96], providing explanations for post removals [50, 51], and placing warning labels in front of inappropriate content [76]. We add to these efforts by taking a user-centered approach and interrogating how victims, offenders and community members experience the current content moderation approach in addressing harm through the perspective of restorative justice.

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 [40, 46, 78, 102–104]. 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 [92] 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 [13]. 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 [16]. Johnson [55] 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 [15]. Yet, restorative justice framework is an especially informative framework for our narrow focus on addressing online harm because of its focus on individuals and relationships rather than content; it therefore forms the focus of our research.

We contribute to 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 [93]. Marwick developed an explanatory model of networked harassment by interviewing people who had experienced harassment and Trust & Safety workers at social platforms [71]. Jhaver et al. conceptualized the distinctions between controversial speech and online harassment by talking to both victims and alleged perpetrators of harassment [52, 53]. Supporting such inquiries, Helberger et al. argues that the realization of public values should be a collective effort of platforms, users, and public institutions [46]. Our research builds on this line of work by considering victims, offenders, and community members’ needs and values in harm reparation to examine the utility of restorative justice approaches in addressing harm.

2.2 Restorative Justice

As the culturally dominant way that we deal with harm, punitive justice is often most familiar to us. Therefore, to explain restorative justice, we will first contrast it with punitive justice. Next, we will detail common practices in restorative justice and its application in offline scenarios.
2.2.1 Restorative Justice Principles. In Western cultures, the dominant model for justice when harm occurs is punishing the offender [105]. The punitive justice model holds that harm is a violation of rules and offenders should suffer in proportion to their offense [36]. 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 [105].

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” [82].

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 [108]. 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 [72]: (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 [72]. Our study locates victims and offenders in the Overwatch gaming community.

2.2.2 Restorative Justice Practices. Restorative justice has been successfully applied in a myriad of settings, such as the criminal justice system, schools and workplaces [108, 116]. When formalized within an organization, restorative justice processes can be embedded within a punitive justice system to use on selected types of harm cases [108, 116]. 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 [4]. A widely used practice in restorative justice is the victim-offender conference [117]. 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 [4, 117]. 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 [117]. 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 [12]. 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.

2.2.3 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 [117]. 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 [25]. 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 [81]. When restorative justice is embedded within a punitive system, the outcome of restorative justice can inform the punishment decision in some cases [108]. 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 [108].

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 [58].

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 [68]. 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 [68]. Restorative justice also requires commitment at the administrative level [38] and time and labor for the parties involved [117].

In this research, we examine the potential of applying restorative justice to online settings for addressing harm. The unique characteristics of online communities—such as anonymity [67], lack of social cues [29], and weak social ties [45]—create new challenges and opportunities for adapting restorative justice practices online. There is a growing interest in the research community to implement restorative justice values to address harm. Blackwell et al. first introduced restorative justice in the content moderation context [9]. 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 [93]. West proposed that education may be more effective than punishment for content moderation at scale [113]. 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 [63]. 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
Addressing harm in online gaming communities

systems, may address this tension [44]. We build on this rich line of work by using the Overwatch gaming community as a case study to explore the benefits and challenges of restorative justice for different stakeholders.

3 BACKGROUND

To provide context for our methods and results, we present a brief overview of the two platforms that we study, Overwatch and Discord.

3.1 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 up 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 the alternative. All players are expected to comply with a set of rules laid out in Blizzard’s code of conduct [11]. For example, these rules instruct that “You may not use language that could be offensive or vulgar 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 [83]. 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 [83]. 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, a temporary account ban, or a 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.

3.2 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.

The moderators in Discord communities are volunteer end-users who regulate their communities and screen posts for inappropriate content [54]. Each community creates its own set of moderation rules. Moderators have the ability to 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 do not have access to private channels, but users can report harm incidents to moderators through private messages. Some communities use automated moderation tools to detect posts containing inappropriate keywords and issue automatic warnings to posters [54]. Prior research has examined how the context of real-time, voice-based communications complicates general moderation practices and poses challenges on Discord [54]. We contribute to the study of Discord moderation by showing how different stakeholders perceive and engage with cases of online harm.

[^3]: https://www.blizzard.com
4 METHODS
We focus on the experiences of three key stakeholders: victims, offenders and moderators. We choose them because they are the key stakeholders in experiencing and addressing online interpersonal harm, and have a corresponding role in a restorative justice conference (victim, offender, and restorative justice facilitator). Thus, focusing on their experiences help us understand the core differences and similarities between punitive or restorative ways of addressing harm. We choose the Overwatch online gaming community as a case study to understand how restorative justice may help address harm. Overwatch is a multi-player game that lets us explore harm in the context of different types of relationships between users, including competition and collaboration. Further, it is considered a “toxic” environment by many players [20, 106]. Like many online communities, the community that we study spans multiple platforms [34]: Overwatch, the gaming platform, and Discord, a discussion platform. We found that Overwatch gamers frequently use Discord to discuss the game, stay connected, and communicate with one another. We also observed that a harm case may initiate in Overwatch but extend to Discord, or vice versa. Therefore, we studied harm cases on Overwatch, Discord, or on both platforms.

To understand gamers’ perspectives on the restorative justice process, we interviewed victims (people who have been harmed), offenders (people who have harmed others)\(^4\), and moderators who dealt with the cases being discussed. We were not able to include Overwatch moderators in our study because they are commercial content moderators [84] who remain anonymous as well as constrained by non-disclosure agreements. We interviewed 23 participants from the Overwatch gaming community for this study.

Restorative justice has largely been developed through practice rather than as an academic discipline. To more deeply understand how its principles might be applied online, we conducted two expert interviews with facilitators from a restorative justice center at a West Coast university. Additionally, the first author attended 30 hours of restorative justice training courses to learn how it is practiced in local communities. These interviews and training enabled us to ground our research in restorative justice values and practices.

4.1 Recruitment
We interviewed a total of 25 participants (Table 1), including 23 participants from three user groups in the Overwatch community: victims, offenders, and Discord moderators. Some participants fall into more than one user group. We recruited participants using a combination of convenience sampling and snowball sampling [7, 85]. 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 for 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

\(^4\)Our use of the terms “victim” and “offender” is for brevity and meant to clarify the role these participants played in specific harm cases we were studying. 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 transformation of the language we use. Using “offenders” may imply that people are “inherently bad” and deserve the condemnation of society [14], while using “victims” may feel disempowering to some and deny the agency victims should have in restorative justice [37]. 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 harm participants, as we found during our data collection. For this early-stage research, we have retained the use of “victims” and “offenders” to keep in line with the language used by our participants.
Discord communities they were involved in. In total, we recruited participants from five Overwatch Discord communities. Besides, 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 recruitment criteria as people who have experienced online harm on Overwatch Discord or the Overwatch game. 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 label themselves with those categories, especially when there has been no opportunity to discuss what has happened [52, 53]. Therefore, we described the recruitment criteria as people who have been warned or banned on Overwatch Discord or in the game. Note that not all of the people who responded to this call considered themselves offenders. During our interviews, some alleged offenders argued that they should not be banned and they were the ones who received harm first (or instead). As researchers, we could not make judgements about the conflicting views between moderators and participants; thus, we analyzed these interviews based on the participants’ views of the harm. This recruitment method did not let us recruit offenders who were not warned or banned. 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.

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., harm that occurred on Discord/Overwatch, harm between friends/strangers, harm within the moderation team/between end-users/between end-users and moderators), and we prioritized participants who had experienced different types of harm for interviews.

4.2 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 [117]. During the first stage of the interviews, we asked questions about the harm case they had experienced, including how it was handled, the impact of harm, and the need to address the harm. During the second stage, we introduced restorative justice principles and the victim-offender conference to the participants and asked questions about their attitudes toward joining the conference to address the harm case they experienced. We chose to focus on victim-offender conference because it is a widely used practice that embeds values from the core restorative justice principles. If time allowed, we asked participants to reflect on more than one harm incident. We noticed that some offenders believed that they had also received harm in that case, or were victims of harm in another case. 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 them about their other self-identified role(s).

Our interviews with Discord moderators were intended to help us understand how they deal with harm in their communities and their attitudes toward using restorative justice to repair the harm. Additionally, since the facilitator is an essential role in offline restorative justice practices, we explored the possibility of creating a corresponding role in online scenarios. Moderators have the closest currently existing role to a facilitator. Therefore, we sought to learn their perspectives on assuming this role. During interviews with moderators, we first asked questions about the harm cases they had handled in their communities and their decision rationales. We then introduced the restorative justice process and asked them to reflect on using it as an alternative approach to
Table 1. Participants' demographic information. We recruited participants using two 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 the self-identified role (victim, offender, moderator, or facilitator; marked by ‘x’) of each in the harm cases they discussed with us. Note that one person may have multiple roles in a single 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        |          |             |
| 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    | Female         | 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        |             |
| P20 | 18    | Male           | Asian     | Some college   | CA     | x        | x        |             |
| P21 | 18    | Male           | White     | Less than high school | US     | x        | x        |             |
| P22 | 20    | Male           | White     | N/A             | UK     | 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.

address the harm cases they had handled in their communities. We also asked moderators for 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 about which they lack previous knowledge. To make restorative justice more concrete, we asked participants to imagine a restorative justice process with the actual harm cases they had experienced or handled. Additionally, we answered participants’ follow-up questions after introducing the restorative justice framework and corrected any misconceptions about it that 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 [18].

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 are handled, as derived from 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 [18]; we still incorporated these interviews because having practitioners analyze
the harm cases victims or offenders experienced can provide valuable insights on how these cases could be alternatively handled in a restorative justice context [91].

We conducted our interviews from February to July 2020. The interviews lasted between one to two hours, and participants received a compensation from $25 to $50 based on the duration of the interview. We conducted 21 interviews using Discord’s voice calls function, while two participants (P1, P13) chose to be interviewed using Discord’s text chat. We also conducted two in-person interviews with the facilitators (P5, P6). 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. Our study was approved by the Institutional Review Board (IRB) at our university.

4.3 Data Analysis

We conducted interpretive data analysis on our interview transcripts [18]. We began with a round of initial coding [89], where we applied short phrases as codes to our data. This first round of coding was done on a line-by-line basis so that the codes stayed close to the data. Examples of first-level codes included “impact of harm,” “creating new account,” “banning,” etc. Next, we conducted focused coding [89] by identifying frequently occurring themes and forming higher-level descriptions. These second-level codes included “notion of justice,” “sense of community,” etc. Coding was done iteratively, where the first author frequently moved between interview transcripts and codes and discussed emergent themes with other authors. After these rounds of initial and focused coding, 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 as we used for other interviews since it helped us compare facilitators’ views of harm to other participants’ views during the analysis.

4.4 Positionality Statement

As researchers working in the sensitive space of understanding online harm and imagining interventions that may help address it, we briefly reflect on our position on this topic. All the authors of this paper feel deeply concerned that online harm is a persistent social problem that disproportionately affects marginalized and vulnerable people [32]. Some authors also come from marginalized groups and are survivors of harm. Our prior research on online harassment has helped us understand the inherent limitations of existing moderation mechanisms; this, in turn, has shaped our desire to look for alternative approaches. We celebrate the growing interest in applying alternative justice theories [6, 48, 61, 73, 112], e.g., racial justice [93] and transformative justice [23], to reduce harm to victims, offenders, communities and societies. We are particularly enthused by the success of restorative justice in addressing offline harm, and its potential in providing agency and care for vulnerable groups who are often ignored and further harmed in a punitive justice model.

While we embrace the values of restorative justice and see its potential, we do not seek to uncritically advocate for restorative justice in this work. All of us live in cultures where punitive justice is the dominant approach of addressing harm, and we have actively worked to learn about restorative justice through research and by practicing its values in our own lives. This education and experience has shaped our interpretations of and understandings of the implications of our findings. At the same time, we acknowledge the limitations inherent in restorative justice processes in addressing deep-rooted, systemic cultural and social issues such as racism and sexism (see transformative justice [23]). Importantly, the authors are aware that restorative justice is voluntary by nature, and forcing the process may cause further harm to the victims. Thus, we have attempted to present an impartial account of our participants’ perspectives on the victim-offender conference and restorative justice principles. While it is not the focus of the current paper, we believe that
practical implementation and evaluation of restorative justice practices in online communities is an important direction for future work to help us more fully understand its utility.

5 CURRENT MODERATION MODELS THROUGH A RESTORATIVE JUSTICE LENS

Restorative justice relies on three main principles which concern victims, offenders and community members: (1) meet the needs of the victims to heal, (2) support the offenders to take accountability and work to repair the harm, and (3) encourage the community to engage in the process to support victims and offenders and heal collectively [72, 117]. We use these three principles to understand the experiences of victims, offenders, and the community during instances of online harm within the current moderation landscape. 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 this content is important to include to help readers understand the nature of online harm.

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

Restorative justice centers the needs of victims. In offline restorative justice processes, victims usually get a chance to share their story, receive emotional support, and provide input on what is needed to repair the harm. The process aims to support victims to heal from the harm and to stop the continuation of harm [117]. In our research, we found that the main tool available to victims to address harm is reporting their offenders to moderators or the moderation system. However, reporting does not effectively address their need for emotional healing or preventing future harm.

To demonstrate the harm participants receive in online gaming communities, we first show an example through case 1.

**Case 1**
P13 teamed up with two 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 she 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
Addressing harm in online gaming communities

5.1 Process and Healing

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 don’t interact with the same offender anymore, incidents of harm are so frequent that they come across new offenders over and over: “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 begin to 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 take conscious efforts to avoid harm in the community, which impacts their experiences in the community. 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 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 as well as lack of transparency and followup inherent in these mechanisms do little to meet victims’ need to heal. In addition, the current reporting mechanism does not effectively reduce offenders in the gaming community because it does not effectively change the culture: even when participants are able to avoid the original offender, they continue to be harmed from new ones.

5.2 Offenders Are Not Supported to Learn the Impact of their Actions or Take Accountability

In a restorative justice view, offenders can address harm through acknowledging their wrongdoing, making reparations, and finally changing their future behaviors. In restorative justice practices such as victim-offender conferencing, offenders learn the impact of their actions through listening to the victim’s side of the story. Afterwards, the offenders can repair the harm and learn through actions such as acknowledging harm (e.g., apologizing), taking anger control courses, or doing community service [25, 117].

Current moderation approaches, as described by our participants, most closely resemble graduated sanctions [64]. 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. In our interviews, we found that many Discord moderators have elements of a restorative mindset: they want to help offenders learn their wrong
doing 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 the consequences of breaking them and can be posted by a moderator or by a pre-configured bot. We find that 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, and 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, the moderator 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 this member 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.’”

The member was finally 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 what oppression is, 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 pronouns tag which demonstrated her gender identity to the community. When she realized this had happened, she swore at the moderators, which led them to permanently ban her from the community.

In case 2, the moderators have made efforts to negotiate with the transgender woman, hoping that she would change by giving her multiple chances and providing explanations for the sanctions. However, the moderation explanations focus exclusively on denying her actions and proving the actions to be against the moderation rules, 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 oftentimes, offenders of harm may have been the victims of harm in other cases. As activist Mariame Kaba puts it: “No one enters violence the first time by committing it.” [42]. 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 have been constantly harmed in gaming communities in the past: “I have been called countless names that are vulgar, offensive, and stuff like that” (P21). Restorative justice believes that although the past harms experienced by offenders do not absolve them of their responsibility for committing 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 [117]. In case 2, the moderators stopped the transgender woman from sharing her story and receiving support. Instead, she received 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 us to work to change the culture besides working with individual offenders. Facilitator P6 explained how a restorative justice approach would strive to provide support and look at the root cause instead of punishing an individual 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 offense, it loses its function when the 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 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 did find evidence of the current punitive approach contributing to maintaining the health of community content and stopping the continuation of harm. Moderators we interviewed told us that some offenders will stop offense after receiving a warning, or will not come back to the same community after being banned. However, when punishment is used as the only means of addressing harm, it can not always effectively do so. We found that punishment often does not help offenders learn the impact of their actions on victims and instead forces them to defend their behaviors. Besides, when it is easy for offenders to avoid punishment (e.g. by creating new accounts), there remain few other mechanisms of accountability.

5.3 Gaming Communities Create Challenges for Victims and Offenders to Address Harm

Restorative justice locates victims and offenders inside a community. Harmful actions disrupt relationships within the community and potentially affect all members. Thus, it is important for community members to acknowledge the harm and participate in redressing it collectively [117]. 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) [72]. In online gaming communities, the micro-community that can come together to address harm can include not only victims, offenders, and moderators, but also bystanders, friends of victims and offenders, and gamers more generally. However, we found that during instances of online harm, victims, offenders and moderators did not come together and had no shared sense of community. As a result, the harm was often considered “someone else’s problem” and remained unaddressed. Also, instead of addressing harm, involvement by micro-community members often created secondary effects that further harmed victims and offenders.
5.3.1 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 towards other gamers who might be harmed by the offenders. 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 feeling related, or further showing care to 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 don’t join with a team. As a result, the offenders don’t need to interact with the victims again 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.”

On many Discord servers, moderators regulate a confined, public space — the general chat. Several moderators told us that 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 and/or ask both parties to resolve problems by 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 make offenders careless about breaking relationships with victims. Additionally, the unclear distinction between public and private online spaces creates challenges for moderators to provide support for victims, and the ensuing lack of support may leave victims vulnerable.

5.3.2 The community creates secondary harm against the victims. In some offline practices of restorative justice conferencing [117], 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 [117]. In our interviews, we found that without this structure and facilitation and care for victims, community members may create secondary harm for the victims. Though it takes courage for victims to share their stories, it can be difficult for them to find an audience that supports and cares about them. 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.

Besides challenging the victims, communities’ ignoring instances of harm can fuel further 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 do 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 do 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 finding suggest that community members are not neutral bystanders when harm happens. They may create secondary harm for victims by ignoring the harm, challenging their stories, or encouraging offenders’ behavior. In contrast, restorative justice encourages the community to build a culture of responsibility and care, where community members collectively prevent and address harm.

5.3.3 The community has a punitive mindset towards offenders. Restorative justice encourages the community to facilitate change in offenders’ behaviors. Community members can share how they have been impacted by the harm, express willingness to support offenders through the change, and acknowledge any changes offenders have made. However, we find that just like the moderation system, the community members who disagree with the offenders’ behavior want to punish them. Case 3 shows an example that happened 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 as a result of 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 left 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, as a 16-year-old, P17 believed the incident changed his career path: “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 to take accountability and change after learning the impact of his actions, besides one gamer friend who reached out to him. This friend suggested that P17 interact more with people offline because he seemed to be 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, the victim P25 had a positive experience when sharing her experiences of harm on Twitter. However, most of the other gamers decided to punish P17 instead of telling him how his actions cause harm. As we mentioned in section 5.2, 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 has apologized to P25, stopped those behaviors, and learned the impact of his action afterwards, he did not have a chance to demonstrate his change to the community. He left the community and abandoned his career goal to be a professional gamer. This runs counter to 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.” It also goes against restorative justice, which believes that ideally, offenders will be welcomed back to the community after their reform [117].

Overall, we found that some community members have a sense of responsibility and care for the victims and for the harm that happens inside the community. However, the way 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 ONLINE RESTORATIVE JUSTICE POSSIBILITIES AND CHALLENGES

In the previous section, we described the current practices to address harm within content moderation systems from the perspective of victims, offenders, and Discord moderators and analyze them through restorative justice principles. We next discuss how these stakeholders responded to the idea of using one common restorative justice practice, namely, a victim-offender conference, to address harm in online gaming communities. Here, we do not seek to analyze participants’ willingness to attend or not attend such conferences as a direct indicator of the potential of online restorative justice. Instead, by asking them to reflect on their needs and concerns with respect to
this practice, we sought to identify the conditions needed for and the challenges of implementing online restorative justice.

6.1 Victims’ Needs and Concerns for a Restorative Justice Process

Many victims in our study expressed a desire to join a victim-offender conference if the offenders are willing to repair the harm. At the same time, many of them had concerns and doubts about the process — including those who showed willingness to attend. We next describe these needs and concerns in detail.

6.1.1 Some victims want to understand and communicate with offenders. Some victims have a need 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 that 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.”

6.1.2 Some victims want an apology, an acknowledgement of mistake, 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 provide them emotional healing. As P13 said, “Just their realization of what was wrong with a small apology would make me happy.”

Several victims hoped that the offenders would change their behavior and stop offending in the future. 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)

6.1.3 Some victims have concerns about whether offenders are willing to repair the harm. During interviews, several victims expressed willingness to join restorative justice meetings if the offenders were willing 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 offenders. These victims did not think that 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 when they consent 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 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 and makes sure that further harm is unlikely to happen before victims and offenders can meet [12].

6.1.4 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 believed that the problems they faced had systemic roots that could not be addressed in a single meeting. P16 is a transgender woman who was harmed by someone whom she believed was transphobic. She 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), was willing to meet with her offender, but she believed that the offender’s aggressive behavior was likely shaped by more than just what happened in the video game. As a result, she believed 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.”

6.1.5 Some victims want to move past the harm. Some victims noted that the harm had already occurred and they wanted to move past it. Several victims were emotionally exhausted by the harm. P20 felt already 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 that 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, we find that the majority of victims have needs other than punishing the offenders. However, they only want offenders to engage with such needs when they are sure that offenders genuinely want to address the harm. In addition, restorative justice approaches might not be sufficient or ideal for online harm that is systemic.
6.2 Offenders’ Interest in a Restorative Justice Process

During the interviews, we applied pre-conference practices in 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. Next, we discuss offenders’ thoughts of the process in detail.

6.2.1 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.”

6.2.2 Some offenders prefer the punitive process over restorative justice. Several offenders (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 felt that they deserved punishment and did not trust that the restorative justice process could 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. The end goal of restorative justice is not to punish offenders, but to help them take accountability of 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. She noted that in school settings, the support for restorative justice from community members, including parents and teachers, encourages the 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)
6.2.3 Some offenders do not fully acknowledge their role in the harm. Several offenders wanted to use the meeting with victims as a chance to go over what had happened in the harm case and 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)

Additionally, as we mentioned in sec. 5.2, the current moderation systems sometimes did not hold offenders accountable for their actions when it’s convenient to use multiple accounts, sometimes anonymously. 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 sample who wanted to repair his relationship with the victim. The mindset of most offenders align with punitive justice. Offenders who acknowledge their wrong-doings think that they deserve punishment and do not believe that restorative justice can lead to a better outcome: on the other hand, offenders who do not fully acknowledge their wrong-doings want to appropriate the victim-offender conference into a punitive process to hold victims accountable. Those views tend to emphasize the consequences the person who cause harm should receive rather than the reparation of harm.

6.3 Moderators’ Views on Implementing Restorative Justice Process in Their Communities

We next discuss Discord moderators’ attitudes towards 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 restorative justice practices. At the same time, they expressed concerns about adapting the moderation process to the restorative justice model. Additionally, several participants had already attempted restorative justice in their communities but received push-back and challenges from moderators and other community members.

6.3.1 There are existing elements of restorative justice in 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 are different. In offline restorative justice practices, facilitators talk with victims or offenders in pre-conferences, where they ask a series of questions to figure out what is needed to repair the harm [81]. Some Discord moderators also talk to victims and offenders before making a moderation decision, but the 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 [12, 108]. On the other hand, the conversation by moderators focuses on facts and evidence, with the moderators acting as judge. 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.”
Besides the process of talking with victims and offenders, both restorative justice and the current moderation system try to understand offender behaviors beyond the current harm case of interest. Restorative justice situates offenders in their life story. 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 [58]. 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.

6.3.2 Moderators’ power may hinder restorative justice process. While we think moderators’ role is the 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 we interviewed also pointed out that while moderators have the final right of interpreting what has happened and who is on the right or wrong, restorative justice seeks to give agency to victims and offenders. Moderators’ power may create prejudices against victims or offenders and reduce their power and 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)

While the removal of power from facilitators is essential, 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 that they have equal chances to express their opinions. [81].
6.3.3 Some moderators think the labor of restorative justice is disproportional to its gains. The pre-conference and conference processes require a significant amount of labor from facilitators [81]. While a facilitator is usually a paid role in offline settings such as schools and prisons [56, 58], Discord moderators are volunteers. Several moderators expressed concerns about the labor required for 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 of them spend hours handling a single harm case. Moderator P14 gave an example: “To ban somebody, we actually 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 not sure 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.”

6.3.4 Some moderators experienced challenges when 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. 6.3.2, P7 conducted a pre-conference with an offender on the Discord server he moderates. P25 is an Overwatch gamer who 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 that it is important to engage 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 establish restorative justice officially as an alternative to the established punitive justice system [56]. They have resources to hire facilitators and can remove or mitigate punishment for offenders who successfully pursue the restorative justice process [58]. Community members also gradually familiarize themselves with restorative justice and provide support to victims and offenders in the process [5, 58].

It is difficult for a community to endorse restorative justice when the established culture and systems are punitive and when there are no examples to 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 is a head moderator in P7’s team, and he 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 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
Table 2. We compare how different actors with a stake in harm are involved in two approaches to addressing harm: a punitive content moderation approach and an approach based on restorative justice. We find that these two approaches differ in their orientations towards victims, offenders, and community members.

|                               | **Punitive content moderation approach** | **Restorative justice approach** |
|-------------------------------|----------------------------------------|---------------------------------|
| **Victims**                   | Victims are left out of the moderation process. | Address victim’s needs for reparation of harm such as support and healing. |
| **Offenders**                 | Offenders receive rule-centered moderation explanations and punishment. | Offenders are encouraged 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. | Engage community members in supporting victims and offenders and heal collectively. |

*new and untested.* Despite such impressions, it is important to examine how we evaluate whether a model for responding to interpersonal harm “works.” For instance, have the victims’ needs been met? Was harm repeated? Do offenders recognize the impact of their actions?

In our interviews, we found that it was hard for people to imagine alternatives to the current moderation system. When we asked participants about their needs beyond having offenders banned or when we introduced restorative justice to them, they often reacted with a pause. They 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)

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

7 DISCUSSION

Content moderation systems predominantly address online harm with a punitive approach. Analyzing through the 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 2 compares the punitive content moderation approach in our sample with a restorative justice approach. As the table shows, restorative justice provides alternative ways to look at what achieving justice means from the perspectives of victims, offenders and community members, which are often absent in current content moderation.

However, applying restorative justice practices in content moderation is not straightforward. Victims, offenders and moderators have needs that are currently unmet and concerns about the process. In this section, we first talk about the challenges of implementing restorative justice. Next,
we offer some possible ways for online communities to implement the approach. Finally, we reflect on the relationship between punitive content moderation approaches and restorative justice.

7.1 Challenges in Implementing Restorative Justice Online

Applying restorative justice practices online is promising, but not straightforward. Though it has the potential to meet the needs of those who have been harmed, there are structural, cultural, and resource-related obstacles to implementing a new approach within the existing punitive framework. Here, we detail potential challenges of implementing restorative justice online. We also discuss some possible ways to address them.

7.1.1 Labor of restorative justice. Restorative justice practices and the process of shifting to them may require significant labor from online communities. Restorative justice conferences can be time intensive for all stakeholders involved. Before victims and offenders can meet, facilitators have to 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 consensus. Stakeholders, in particular facilitators, must expend emotional as well as physical labor. In offline practices, facilitators usually occupy a paid role in organizations [38]. 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 physical and emotional labor with a growing community [30, 83, 100]. 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 to receive 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 above mentioned 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 some precedent for platforms making such investments—in recent years, social media companies such as Discord have hosted trainings for its community moderators 5. Meanwhile, a bottom-up process could engage users with preexisting knowledge about restorative justice to first introduce the process in their communities and then to gradually expand the restorative culture and practice from there. In our sample, two moderators have attempted or practiced online restorative justice within their own communities. They showed enthusiasm for expanding restorative justice practice in other online gaming communities. It is possible that resources from companies and practitioners could collectively begin to address the labor problem.

An additional consideration here is that the implementation of online restorative justice requires a re-allocation of labor instead of merely an addition of labor. We found that many moderators we interviewed already practiced different elements of restorative justice. Some moderators aim to support victims and give 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 asks for a shift of purpose in existing processes – from determining point of offense to caring for victims and offenders. In addition, if online restorative justice can stop the perpetuation of harm more frequently than punitive justice, it can reduce the need for moderation labor in the long term. While research has shown that offline restorative justice has successfully reduced re-offense rates [25, 68, 75], how effective restorative justice can be in online communities is an important area for future work.

---

5Discord Moderator Academy. https://discord.com/moderation
7.1.2 Individuals’ understanding of justice align with the existing punitive justice model. We found that although people have needs that the current system does not address, their mindsets and understandings of potential options are often aligned 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 because of its widespread application throughout society, including in prisons, workplaces, and schools [35, 62].

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 were reluctant to engage in restorative justice because they still expected to receive bans after the process, but restorative justice usually serves as an alternative to instead of an addition to punishment. 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 need to be aware of them. We believe that one crucial step toward applying alternative justice frameworks (e.g. restorative justice) online is to provide people with information and education on restorative justice. Helping people understand the diversity of restorative justice processes and how its aim is restoration instead of punishment may address their doubts and open up more opportunities for change. Besides, 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 evolved in practice [24, 117], and it is important for people to experience it themselves in order 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 that can be adapted to online communities. In future work, we hope to pilot online restorative justice sessions and run survey studies to understand the types of harm that are more likely to benefit from restorative justice and the types of processes that would be useful. We would also like to provide more precise empirical guidelines about how restorative justice can be embedded in moderation systems. To manifest a future of restorative justice, “we will figure it out by working to get there” [57].

7.1.3 Offender accountability. Another challenge may be in motivating offenders to take accountability for their wrong-doing – 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 without clear rules of moderation, e.g., if harm occurs during a private chat in Discord, or happens across multiple platforms. Some participants also noted that it is easy to have multiple identities in Overwatch or Discord. Thus, when punishment is not effective, punitive justice may also lose its effectiveness.

Punishment is not the only form of holding offenders accountable. Restorative justice believes that people are also tied through relationships. Our interview data as well as restorative justice literature suggest the importance of a sense of community—if offenders perceive themselves to be members of a shared community with victims, they will be more likely to take accountability
to address harm \cite{58, 81}. 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 places 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.

7.1.4 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, which prohibits 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 \cite{101}. Face-to-face sharing enables victims and offenders to see each other’s body language and facial expressions. Besides, offline restorative justice has 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 \cite{81}. 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.

When disingenuity exists, its costs can be high. Some restorative justice processes, e.g., issuing just an apology, are low-effort and could be open to exploitation by disingenuous offenders to escape punishment. However, other more complex processes, such as the victim-offender conference requires time and effort on the part of offenders to acknowledge their wrong doing and to execute actions that repair the harm, which is not a necessity in punitive justice \cite{117}. Indeed, prior research as well as our findings suggest that offenders do not see restorative actions to repair harm (e.g., community service) as an easier option compared to punitive processes \cite{114}. Thus, more involved restorative processes could be more effective at restoring offenders.

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 are being developed to facilitate civil engagement and communication in online systems. For example, Kriplean et al. built a platform, Considerlt, to support reflective interpersonal deliberation \cite{65}. REASON (Rapid Evidence Aggregation Supporting Optimal Negotiation) is a Java applet developed to improve information pooling and arrive at consensus decision \cite{49}. Many researchers have attempted to model human-like characteristics and emotional awareness in chatbots \cite{1, 99}. In the context of restorative justice approach, Hughes has developed a tool, Keeper, for implementing online restorative justice circles \cite{47}. Such existing systems can be leveraged to develop advanced tools that facilitate emotion-sharing and communication in online restorative justice processes.

7.2 Applying Restorative Justice in Online Moderation

In this section, we discuss some possible ways to adapt current moderation practices and implement restorative justice in moderation to address online harm. 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 go beyond the conference. Indeed, restorative justice is not a fixed set of values or practices, but evolves over time and across contexts. Implementing restorative justice practices online would require us to adapt them to meet the needs
and limitations of the online environment. We suggest that online communities can start with partial restorative justice practices that only involve some stakeholders (e.g. pre-conferencing), or adapt from some of the existing moderation practices to embed restorative justice values, and implement victim-offender conferencing when the preconditions are met.

7.2.1 Embed restorative justice language in moderation explanations. Moderators can embed restorative justice language in explanations of moderation decisions. Prior work has shown that providing explanations for moderation decisions can reduce re-offense in the same community, but many moderated offenders also dismiss such explanations and continue re-offending [50, 51]. In our research, we find that explanations often focus on describing community guidelines or presenting community’s norms to the users, but do not encourage offenders to reflect on the impact of their actions. These explanation 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 for 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.”

Additionally, victims are often left out of the moderation process in the online communities we studied. However, some victims want to get information on the moderation process and result. While prior research has talked about how moderation transparency prevents offenders from re-offending [51], our work points to the importance of providing information for victims in the moderation process. Besides, victims have needs for support and healing. We suggest that a note of care may help victims feel heard and validated and help them recover.

7.2.2 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 are with offenders or victims alone [66, 72]. Some Discord moderators already talk with victims and offenders before making moderation decisions, a practice redolent of restorative justice pre-conferencing. In offline pre-conferencing, facilitators ask a series of questions to help victims and offenders reflect on the harm, which provides opportunities for healing and learning to them. We propose that pre-conferencing provides opportunities to meet some of the needs our participants identified. For example, some offenders want to issue an apology to the victims; victims may not want to meet offenders but just let their feelings be known through the facilitator.

In offline restorative justice, pre-conferencing is a preparation step for a potential victim-offender conference. Thus, if both the victim and the offender indicate willingness to meet with each other in the pre-conference, the moderators can organize a victim-offender conference, where victims and offenders 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 [12]. 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 [27]. In our sample, many victims indicated a need for receiving emotional support. We propose that online communities can provide a space for victims to share the harm they experienced and receive support. Systems with similar goals have previously been built in the online spaces. For example, Hollaback is a platform that allows victims of street harassment to empower each other through storytelling [28]. Heartmob enables victims to describe their experiences of harm and solicit advice and emotional support from volunteer Heartmobers [10]. 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 the community members to provide support.

7.3 Situating restorative justice in the moderation landscape

We have so far illustrated some possible ways to implement restorative justice in the current moderation system. Yet, we do not seek to advocate for restorative justice as a whole 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.

We propose that restorative justice conferencing should only be used in select cases because it is only effective when it is voluntary and the parties involved are committed to engaging in it. Our findings show that individuals have different conceptualizations of justice. As Schoenebeck et al. also found, people’s preferences of punitive or restorative outcomes vary with their identity or social media behaviors [93]. 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 [38, 108].

We estimate that whether an online community decides to apply restorative justice is a value related question. While restorative justice and punitive justice processes have the shared goal of addressing harm and preventing the perpetuation of harm, they have different processes and orientations towards when justice is achieved. Content moderation—closer to a punitive justice approach—addresses harm through placing limits on the types of content that remain on the site and punishing offenders in proportion to their offense. Restorative justice, on the other hand, 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 how it helps to achieve the current goals of effectively removing inappropriate content and punishing offenders, but how online communities may benefit from restorative justice values and goals. Seering et al. found that community moderators have diverse values in what moderation should achieve [95]. 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, it is important for online communities to reflect on their values and goals and decide on what mechanisms (e.g. whether punitive or restorative justice) help realize those values. Recent research has argued that social media platforms have a responsibility to incorporate ethical values in moderation instead of merely optimizing to achieve commercial goals [46]. In particular, some researchers have proposed values and goals that align with restorative justice, such as centering victims’ needs in addressing harm [10, 93], democracy [94], and education [113]. 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, we argue that 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 the offenders [3]. 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, communities may focus on how victims, offenders and the entire community could benefit from the process.

7.4 Limitations and Future Work

Our study has some methodological limitations. We used convenience and snowball sampling to recruit participants in one gaming community. As a result, our participants’ experiences are not representative of all types of harm online. While Overwatch has a multinational and multicultural userbase, most of our participants are English-speakers from western cultures. In future work, we would like to conduct large-scale surveys to understand how the characteristics of harm influence peoples’ needs for a restorative justice process in addressing online harm.

There are diverse values and practices in restorative justice all over the world and they are evolving rapidly. We believe that future research can build on and expand restorative justice beyond the three principles and the victim-offender conference. In addition, while our research motivates a need for restorative justice, future research is needed to see how restorative justice can be integrated into an online community. In future work, we plan to collaborate with online communities to implement and test restorative justice practices.

While this research focused on the restorative justice approach, it is not the only alternative and has its limitations. Expanding people’s notion and understanding of justice creates opportunities to utilize the strengths of different justice frameworks. As some of our participants mentioned, many harms are rooted in structural issues such as sexism and racism. Effectively addressing such issues would require a deep, collective commitment to transforming the underlying conditions that enable harm to occur. Transformative justice [23] and community accountability [74] are some of the frameworks of justice that have been developed for this purpose. We believe that the work of transforming the underlying conditions that foster harm must simultaneously be done by everyone and in all communities for a sustained change to occur.

8 CONCLUSION

In this research, we conducted interviews with victims, offenders and moderators in the Overwatch gaming community. We 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 for exploring what restorative justice may look like in an online community and when and how to implement this approach. We hope that this work serves as a useful guide for designers, volunteers, activists and other scholars to examine a restorative justice approach to addressing online harm.

REFERENCES

[1] Achini Adikari, Daswin De Silva, Damminda Alahakoon, and Xinghuo Yu. 2019. A cognitive model for emotion awareness in industrial Chatbots. In 2019 IEEE 17th international conference on industrial informatics (INDIN), Vol. 1. IEEE, 183–186.

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31, 2 (may 2017), 211–236. https://doi.org/10.1257/jep.31.2.211

J. ACM, Vol. 37, No. 4, Article 111. Publication date: November 2022.
[3] Gordon Bazemore and Mara Schiff. 2015. *Restorative community justice: Repairing harm and transforming communities*. Routledge.

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

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

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

[7] Patrick Biernacki and Dan Waldorf. 1981. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research* 10, 2 (Nov. 1981), 141–163. https://doi.org/10.1177/004912481810002005 Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[8] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*. Springer, 405–415.

[9] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *Twelfth International AAAI Conference on Web and Social Media*.

[10] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.

[11] Blizzard. 2020. Blizzard’s In-Game Code of Conduct. https://us.battle.net/support/en/article/42673

[12] Jane Bolitho and Jasmine Bruce. 2017. Science, art and alchemy: Best practice in facilitating restorative justice. *Contemporary Justice Review* 20, 3 (2017), 336–362.

[13] Shannon A Bowen. 2013. Using classic social media cases to distill ethical guidelines for digital engagement. *Journal of Mass Media Ethics* 28, 2 (2013), 119–133.

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

[15] Caitlin Ring Carlson and Luc S Cousineau. 2020. Are You Sure You Want to View This Community? Exploring the Ethics of Reddit’s Quarantine Practice. *Journal of Media Ethics* 35, 4 (2020), 202–213.

[16] Archie B Carroll. 2016. Carroll’s pyramid of CSR: taking another look. *International journal of corporate social responsibility* 1, 1 (2016), 1–8.

[17] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet’s Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

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

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

[20] Chunia. 2019. This community is toxic [Online forum post]. https://us.forums.blizzard.com/en/overwatch/t/this-community-is-toxic/290489

[21] Danielle Keats Citron. 2015. Addressing Cyber Harassment: An Overview of Hate Crimes in Cyberspace. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2932358

[22] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing Revenge Porn. *Wake Forest Law Review* 49 (2014). http://heinonline.org/HOL/Page?handle=hein.journals/wflr49[&]id=357[&]div=15[&]collection=journals

[23] Erin Daly. 2001. Transformative justice: Charting a path to reconciliation. *Int’l Legal Persp.* 12 (2001), 73.

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

[25] Kathleen Daly. 2002. *Restorative justice: The real story*. *Punishment & Society* 4, 1 (Jan. 2002), 55–79. https://doi.org/10.1177/1462474022228464

[26] Lina Dencik, Arne Hintz, and Jonathan Cable. 2017. Towards data justice. *DATA POLITICS* (2017), 167.

[27] James Dignan. 2001. *Understanding victims and restorative justice*. McGraw-Hill Education (UK).

[28] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. 2013. Hollaback! The role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 477–490.

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

[30] Bryan Dosono and Bryan Semaan. 2019. 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*. 1–13.

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

[32] Maeve Duggan. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[33] Maeve Duggan and Aaron Smith. 2013. 6% of online adults are reddit users. (2013).
Addressing harm in online gaming communities

[34] Casey Fiesler and Brianna Dym. 2020. Moving Across Lands: Online Platform Migration in Fandom Communities. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 042 (May 2020), 25 pages. https://doi.org/10.1145/3392847

[35] Michel Foucault. 2012. Discipline and punish: The birth of the prison. Vintage.

[36] David Garland. 2012. Punishment and modern society: A study in social theory. University of Chicago Press.

[37] Theo Gavrielides. 2017. Collapsing the labels ‘victim’and ‘offender’in the Victims’ Directive and the paradox of restorative justice. Restorative Justice 5, 3 (2017), 368–381.

[38] Theo Gavrielides. 2017. Restorative justice: Ideals and realities. Routledge.

[39] Tarleton Gillespie. 2018. Custodians of the Internet | Yale University Press. https://yalebooks.yale.edu/book/9780300173130/custodians-internet

[40] Robert Gorwa. 2019. What is platform governance? Information, Communication & Society (2019), 1–18.

[41] Joshua E Introne. 2008. Adaptive mediation in groupware. Brandeis University.

[42] Jon Hurwitz and Mark Peffley. 2005. Explaining the great racial divide: Perceptions of fairness in the US criminal justice system. The Journal of Politics 67, 3 (2005), 762–783.

[43] Michel Foucault. 2012. Punishment and modern society: A study in social theory. University of Chicago Press.

[44] Amy A Hasinoff and Nathan Schneider. 2022. From Scalability to Subsidiarity in Addressing Online Harm. Social Media+ Society 8, 3 (2022), 20563051221126041.

[45] Caroline Haythornthwaite. 2002. Strong, weak, and latent ties and the impact of new media. The information society 18, 5 (2002), 385–401.

[46] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. The information society 34, 1 (2018), 1–14.

[47] Maggie Hughes and Deb Roy. 2020. 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. 275–279.

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

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

[50] Amy A Hasinoff and Nathan Schneider. 2022. From Scalability to Subsidiarity in Addressing Online Harm. Social Media+ Society 8, 3 (2022), 20563051221126041.

[51] Theo Gavrielides. 2017. Collapsing the labels ‘victim’and ‘offender’in the Victims’ Directive and the paradox of restorative justice. Restorative Justice 5, 3 (2017), 368–381.

[52] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294

[53] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. ACM Trans. Comput.-Hum. Interact. 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[54] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–23. Publisher: ACM New York, NY, USA.

[55] Brett Gregory Johnson. 2017. Speech, harm, and the duties of digital intermediaries: Conceptualizing platform ethics. Journal of Media Ethics 32, 1 (2017), 16–27.

[56] Gerry Johnstone and Daniel Van Ness. 2013. Handbook of Restorative Justice. Routledge. Google-Books-ID: U2UQBAAABBQAJ.

[57] Marianne Kaba. 2021. We Do This Til We Free Us: Abolitionist Organizing and Transforming Justice. Haymarket Books.

[58] Charles Kiene and Benjamin Makoto Hill. 2020. 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. 1–8.

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

[60] Mimi E. Kim. 2018. From carceral feminism to transformative justice: Women-of-color feminism and alternatives to incarceration. Journal of Ethnic & Cultural Diversity in Social Work 27, 3 (2018), 219–233. https://doi.org/10.1080/15313204.2018.1474827 arXiv:https://doi.org/10.1080/15313204.2018.1474827
[91] Niloufar Salehi, Jaime Teeven, Shamsi Iqbal, and Ece Kamar. 2017. Communicating context to the crowd for complex writing tasks. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 1890–1901.

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

[93] Sarita Schoenefbeck, Oliver I. Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. New Media & Society 2020, 1–28.

[94] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2020. Metaphors in moderation. New Media & Society 2020, 1–28.

[95] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. 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. 111–125.

[96] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. New Media & Society 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[97] Orit Shaer, Lauren Westendorf, Nicholas A Knouf, and Claudia Pederson. 2017. Understanding gaming perceptions and experiences in a women’s college community. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 1544–1557.

[98] Timo Spring, Jacky Casas, Karl Daher, Elena Mugellini, and Omar Abou Khaled. 2019. Empathic Response Generation in Chatbots. In SwissText.

[100] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445092

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

[102] Nicolas Suzor. 2018. Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. Social Media + Society 4, 3 (2018), 2056305118787812.

[103] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. International Communication Gazette 80, 4 (2018), 385–400.

[104] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. International Journal of Communication 13 (2019), 18.

[105]Zenon Szablowinski. 2008. Punitive justice and restorative justice as social reconciliation. The Heythrop Journal 49, 3 (2008), 405–422.

[106] Vampragon43. 2020. Why Overwatch is so toxic [Online forum post]. https://us.forums.blizzard.com/en/overwatch/t/why-overwatch-is-so-toxic/445652

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

[108] Daniel W Van Ness. 2016. An overview of restorative justice around the world. (2016).

[109] Jonathan Vanian. 2017. Twitter Toughens Rules on Nudity and Revenge Porn | Fortune. http://fortune.com/2017/10/27/nudity-revenge-porn-twitter/

[110] Emily A Vogels. 2021. The state of online harassment. Pew Research Center 13 (2021).

[111] Charlie Warzel. 2016. A Honeypot for Assholes: Inside Twitter's 10-Year Failure to Stop Harassment. BuzzFeed News (2016).

[112] Michael Wenzel, Tyler G Okimoto, Norman T Feather, and Michael J Platow. 2008. Retributive and restorative justice. Law and human behavior 32, 5 (2008), 375–389.

[113] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society 2018.

[114] Eric J Wodahl, Robbin Ogle, Colleen Kadlec, and Kenneth Gerow. 2013. Offender perceptions of graduated sanctions. Crime & Delinquency 59, 8 (2013), 1185–1210.

[115] Donghee Yvette Wohn. 2019. 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. 1–13.

[116] William R Wood and Masahiro Suzuki. 2016. Four challenges in the future of restorative justice. Victims & Offenders 11, 1 (2016), 149–172.

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