Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games

Renkai Ma

The Pennsylvania State University, renkai@psu.edu

Yao Li

The University of Central Florida, yao.li@ucf.edu

Yubo Kou

The Pennsylvania State University, yubokou@psu.edu

Multiplayer online games seek to address toxic behaviors such as trolling and griefing through behavior moderation, where penalties such as chat restriction or account suspension are issued against toxic players in the hope that punishments create a teachable moment for punished players to reflect and improve future behavior. While punishments impact player experience (PX) in profound ways, little is known regarding how players experience behavior moderation. In this study, we conducted a survey of 291 players to understand their experiences with punishments in online multiplayer games. Through several statistical analyses, we found that moderation explanation plays a critical role in improving players' perceived transparency and fairness of moderation; and these perceptions significantly affect what players do after punishments. We discuss moderation experience as an important facet of PX, bridge the game and moderation literature, and provide design implications for behavior moderation in multiplayer online games.

CCS CONCEPTS • **Human-centered computing • Human computer interaction**

Additional Keywords and Phrases: Multiplayer online games; behavior moderation; toxicity; moderation design

ACM Reference Format:

Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23--28, 2023, Hamburg, Germany. https://doi.org/10.1145/3544548.3581097

1 INTRODUCTION

Multiplayer online games face rampant toxic behaviors such as trolling, harassment, and griefing in their player communities [1,17,28,59,61], and game researchers and practitioners have long recognized the grim challenge of designing effective moderation systems that discourage toxic behaviors while encourage cooperative ones (e.g., [31,64,76,90]). Game platforms also routinely update their moderation policy and practice to cover new types of toxicity [70,71]. Their present moderation systems usually adopt a punitive model, where players, once convicted, will receive one or more penalties, ranging from losing access to certain in-game privileges (such as in-game rewards) to permanent account suspension. Game companies may issue penalties against individual players or a group of players for unwanted behaviors (e.g., [7,82]). Numerous celebrity players make headlines for receiving permanent bans for their in-game toxicity (e.g., [3,6,30,86]).

Given how commonplace it is for game companies' moderation systems to hand out penalties to players, understanding players' moderation experience is of important value to multiplayer online games for many reasons. First, although moderation plays a central role in managing toxicity in multiplayer online games [76], there is limited understanding regarding its effectiveness in reforming player behavior if those punished players are to stay. Second, receiving a moderation penalty intersects deeply with the player experience (PX) in game if they are temporarily banned from play. Third, a moderation penalty could often incur negative emotions such as frustration and anger [26,98], thus intersecting with players' emotional experiences.

The punitive model of online moderation systems is not without limitations when a penalty is likely the first and only point of contact between users and the moderation system. In other words, the design of punishment matters. Prior moderation research, most of which is done in the context of social media platforms such as Reddit [45,48], Facebook [73,91,98], and Instagram [27,35], has provided ample reflections on this. For instance, users may not understand why they are punished and have to figure that out on their own [48,57]. Affected users need more fairness, accountability, and transparency in punishment design in order to develop better trust [47,81,98,107]. Rich empirical findings from the recent moderation literature suggest that how users experience punishments matters to their compliance with behavioral standards as well as their later conducts [48,51]. However, little attention has been paid to how players experience moderation in multiplayer online games. In this study, we use behavior moderation and punishment design interchangeably, where behavior moderation is more conceptual and denotes a cluster of approaches to manage player behavior, while punishment design is more operational and represents specific moderation actions that players experience.

To approach this question, we conducted a survey in May 2022 to understand how players in multiplayer online games experience punishments from behavior moderation. Specifically, the study leverages the existing moderation literature (e.g., [44,47,66,98,99]) to focus on the perceived transparency and fairness of moderation as well as the intended adoption of coping strategies for punishments in the context of online gaming. We performed exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and structural equation modeling (SEM) on the valid survey data (N=290). We found that while punishment notification and explanation significantly increase players' perceived transparency of moderation, explanation provision plays a more critical role than explanation and punishment types in increasing all notions of fairness perceptions. Also, as perceived fairness, especially retributive, restorative, and procedural justice, plays more critical roles than perceived transparency in motivating players' intended adoption of coping strategies for punishments, moderation explanation as one punishment design became more important to help punished players cope with punishments. We discuss how these findings extend our understanding from the moderation literature that primarily focuses on the social media context (e.g., [44,47,51,67,98]). We then discuss the necessity of considering players' moderation experience as part of PX and derive practical implications for moderation design and policymaking in game from our findings.

We contribute to HCI and game research in four ways: First, we contribute quantitative insights into players' moderation experiences. Second, we contribute survey items for assessing players' punishment/moderation experiences with high validity and reliability for future work that focuses on this topic. Third, we theorize players' moderation experiences in relation to player experience by bridging the moderation literature and the player experience literature. Lastly, we contribute concrete design implications for moderation in multiplayer online games.

2 BACKGROUND: TRANSPARENCY, FAIRNESS, AND COPING WITH MODERATION PUNISHMENT

In line with rising ethical concerns about algorithmic systems (e.g., [20,63,65]), HCI researchers have recently paid attention to transparency and fairness in users' experiences with moderation systems (e.g., [47,68,98]). Transparency implies openness and communication [93], allowing users to "uncover the true essence of a system" [15]. A large body of prior work has seen moderation notification and explanation as important design approaches for users to understand moderation system's decision-making (e.g., [47,56,67,98]). Moderation notifications and explanations as platforms' transparency efforts thus are critical for users to assess moderation transparency. Fairness can be defined on diverse ontological bases. Moderation researchers have initiated various discussions by leveraging diverse dimensions of fairness notions, such as procedural or restorative justice, to assess moderation fairness (e.g., [67,85,104]).

Investigating the perceived transparency of moderation systems is a growing research interest. When Facebook failed to inform users of content removal at the time of its issuance [91], users questioned what content rules Facebook deemed they violated [73]. Researchers also uncovered that users complained about the inconsistent punishments that happened between them and others, and thus the users requested further explanations (e.g., [68,98]). Prior work also stressed the importance of disclosing sufficient information in moderation explanations [48], which could be educational to punished users for behavior reform [45] and build up trust for platforms [91]. Especially, as harmful content can be categorized based on different severities [84], it becomes important for platforms to make transparency efforts pertaining to the varying severities to show how a moderation decision is made.

Beyond the moderated users' perspectives, transparency is also a design consideration stressed by human moderators who practice moderation in game-related contexts (e.g., live-streaming platforms). Sometimes, it could be intuitive that human moderators inform rule breakers what and why they are accused. For example, Cai et al. found that volunteer moderators on the live streaming platform, Twitch, actively communicate with rule violators to ensure moderation practices and decisions are in an appropriate degree of visibility to the public [12]. But oftentimes, it is challenging for platforms to decide to disclose what degree of moderation transparency. Jiang et al. found that on Discord, human moderators usually encounter challenges in making all content rules explicit and transparent because user behaviors are complex and nuanced in voice-based communities [51]. Kou and Gui found that gamers who flag and report other players doubt the transparency of flagging mechanism, especially around whether and how it works, so gamers generate distrust to moderation system [59]. But still, researchers have generally reached a consensus that moderators should keep different moderation procedures transparent such as moderation notification and explanations [45,51,73], as well as appeal process [27,52].

Beyond moderation transparency from the angles of either moderated users or human moderators, researchers have also paid attention to the fairness perception of moderation system. Since fairness has scarcely been defined in a consensus, many researchers have used multiple dimensions of fairness, such as *outcome fairness*, *retributive*, *procedural*, and *restorative justice*, to understand users' perceived fairness of moderation system. First, *outcome fairness* means the extent to which users perceive the distribution of moderation decisions (e.g., account suspension [98], visibility deduction [68]) is fair. Several prior studies have found that users would perceive content removal or account suspension as fair if they receive a moderation explanation (e.g., [44,98]). Second, *procedural justice* refers to fair processes where users' fairness perceptions are influenced by their experiences [108]. A recent study has found that content creators experienced inconsistent punishments that simultaneously violated the platform's content rules, so creators felt moderation as unfair [68]. Third, *retributive justice* describes correct justice processes,

where people violating rules require to suffer proportionally in return [101]. However, many researchers have been concerned that retributive justice might not be the only effective justice standard for adjudicating moderation cases (e.g., [8,57]). That is because users might not be able to effectively learn from what they have done wrong by server punishments but continue being toxic [44]. Thus, one of the alternative justice models, *restorative justice*, is appropriate to re-assess moderation cases. This justice notion seeks to have both offenders and victims in the justice process and allow their voice to be heard by decision-makers [101]. That means, platforms need to communicate with punished users [67], or as a recent study suggested, there at least should be community efforts involving punished users, victims, and other community members to justify moderation cases together [104]. Given these four dimensions of fairness notion from prior work, we position our study in an integrative fashion of involving them to interpret players' perceived fairness of behavior moderation in game. This can enable us to uncover more nuances that might be missed if we use one single notion of fairness.

Along with studying user perceptions of moderation, researchers have started to investigate how users cope with punishments (e.g., [26,67,73]). Coping is "the person's constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the person's resources" [28]. Prior moderation literature has discussed how users might or might not have enough resources to handle punishments behaviorally or cognitively. For example, users are found to avoid future punishments by tweaking their content [14,27,36] or creating closed, hidden groups to connect with and support other punished users [35]. Users were also found to resist moderation punishments by generating memes publicly to express complaints [92] or initiating appeal procedures (e.g., [68,98]). Besides such behavioral efforts in responding to punishments, users also choose to make cognitive efforts, such as getting recovered and healing through communication with other community members [27] or collectively conducting sense-making on why punishments happen [67]. When users do not have enough effort to behaviorally cope with punishments, they choose to pay little attention to punishments and accept what impacts the punishments bring to them [11].

So, to better understand users' coping efforts for behavior moderation and punishments in the context of online gaming, we borrowed the five *coping* strategies delineated by Scherer et al., including (1) *problem-focused coping*, (2) *detachment,* (3) *wishful thinking*, (4) *seeking social support*, and (5) *focusing on the positive* [83]. *Problem-focused coping* means that people come up with different solutions for the problem; *detachment* refers to the strategized attitude of seeing the problem as nothing happening; *wishful thinking* means that people wish the problem would go away or a miracle will happen; *seeking social support* refers to talking to or asking support from someone about the problem; *focusing on the positive* denotes peoples' actions of redirecting attention from the problem to something positive or creative.

2.1 Research Gap: Moderation/Punishment Experience in Game

A growing body of research has examined users' moderation experiences in game-related contexts (e.g., live streaming platforms and audio-based community). Importantly, these game-related contexts are a unique type of online community in nature, and thus fundamentally different from the game contexts in several ways. These users, like many social media users, might complain moderation decision-making as opaque or unfair (e.g., [26,73,91]), but their behaviors are mostly presented as user-generated content (e.g., videos [69] on YouTube, textual content on Reddit [45,48], or audio on Discord [51]) in a relatively static way. That means, either human moderators or moderation algorithms could find ways (e.g., hash/keyword matching or classification [39]) to identify and legitimately adjudicate whether they violate content policies [12,45,79], and moderation decision-makers could

further notify what they have accused users for [47,98]. However, game players' behaviors are categorically more complex through a combination of in-game communication, avatar actions, interactions with game design, etc. Problematic user content such as hate speech that is commonly found on social media is just one type of violation for moderation in game; and players might commit toxic behaviors that exclusively happen in online multiplayer games, such as sabotaging teamwork or intentionally getting killed by enemies [66]. Or even the game platform designs, such as matchmaking system or players' perceived loss due to merit-based game competition, could be part of the reason that players become toxic and violate platform policies. As such, player behavior is notoriously more difficult to adjudicate than social media users' behaviors which are mostly text, audio, or video-based.

All of the aforementioned concerns introduce profound challenges to moderation fairness and, subsequently, players' fairness perception in multiplayer online games. Unsurprisingly, multiplayer online games have long wrestled with what constitutes a fair moderation decision. As early as the 1990s, MUD users would debate what penalty was proper for a user who had committed a virtual "rape" [19]. In contemporary multiplayer online games, game companies such as Riot Games must deal with their player base' reactions to permanent bans of gaming celebrities [40]. Also, a great problem fronted by both researchers and gaming companies is how to design better moderation systems to help reform player behaviors. Researchers who have focused on social media or game-related moderation have thought about rethinking moderation decision-making procedures (e.g., involving users' voice for procedural justice [26] or expert review in moderation [79]) or designing moderation explanations that can instruct users about content policies [48,52].

However, given the complexity in player behaviors, it would be hard to come up with design solutions to improve in-game moderation system unless we advance the understanding of how players experience punishments. Calling for more attention to players' moderation experiences, we recognize the nuances of in-game punishment design, compared to moderation in other contexts such as game-related or social media communities. For example, players might receive ranked rating deduction (i.e., game skill level decrease) or barred entry from joining certain types of game (e.g., matchmaking or queue restrictions) from competitive games. While game-related contexts such as communities on Twitch or Discord conduct similar moderation mechanisms (e.g., chat restriction [87]), they might not be enough or contextual for the in-game environment, which are usually competitive, merit-based for winning, and toxic [56,88]. Thus, to reflect on and implicate better punishment design in game, we aim to fill the research gap of players' moderation experiences - their perceptions of and coping reactions to punishment.

3 RESEARCH QUESTIONS & HYPOTHESIS DEVELOPMENT

This section will discuss how we distill two specific research questions (RQ1 and RQ2) and corresponding hypotheses within two hypothesized models from the prior work. Our first hypothesized model (H1-H3) for RQ1 describes the purposed relationships between punishment design (e.g., notification, explanation) and punished players' perceptions of moderation (i.e., perceived transparency and fairness) as well as their intended adoption of coping strategies for punishments, as summarized in Figure 1. Figure 2 summarizes the purposed relationships (H4) for RQ2 between the perception of moderation, including perceived fairness and transparency, and the intended adoption of coping strategies.

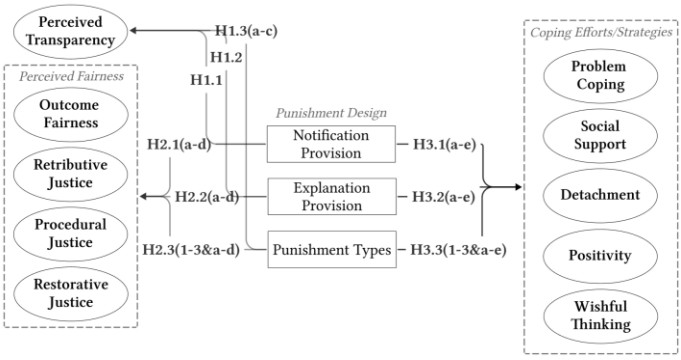

Figure 1 Hypothesized model 1 of punishment experience for RQ1 (H1-H3). We removed "wishful thinking" after exploratory factor analysis since its all survey items had significant cross-loadings on other factors (see Section 5.1).

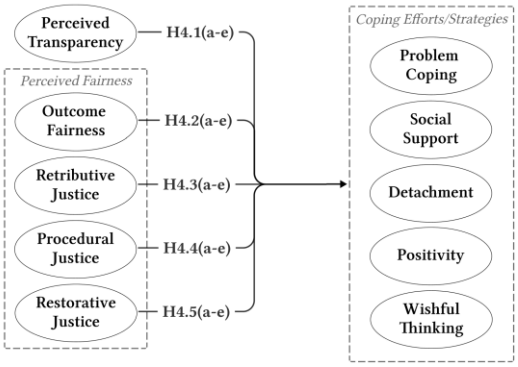

Figure 2 Hypothesized model 2 of punishment experience for RQ2 (H4). We removed "wishful thinking" after exploratory factor analysis since its all survey items had significant cross-loadings on other factors (see Section 5.1).

3.1 Punishment Design

Punishment design means the design construction where players experience punishments. Prior work has broadly understood moderation notification and explanation as important design components or facets in moderation processes (e.g., [47,56,67,98]). For example, in the issuance of moderation punishments such as account suspension, Suzor et al. found users felt confused about punishments since they did not receive notifications [91]. When users try to make sense out of punishments, they request detailed explanations of what policy they were deemed to violate by moderation system [73]. In the context of online gaming, such explanation can be detailed as reasons for punishment (e.g., "reform card" in League of Legends [58]), resource provision to cope with punishment (e.g., information on how to appeal punishments [98]), and more to help punished players understand moderation decision-making. When punished users enter an appeal procedure of punishment, Vaccaro stressed the importance of explanation provision to improve their perceived fairness and trustworthiness of platform [98]. Also, throughout moderation processes like receiving and appealing punishments, game platforms might utilize different punishments to govern different identities of users, such as professional players, coaches, and teams [66]. Thus, punishment notification, explanation, and named punishments are three punishment design components important to punished users and also shared by platforms. As these punishment design components shape users' moderation

experiences, we aim to understand how they affect players' perceived transparency and fairness, as well as the intended adoption of coping strategies for punishments, as we discussed in Section 2. So, we ask:

RQ1 Does punishment design affect players' perceptions of behavior moderation and their post-moderation behaviors?

RQ1.1 Does punishment design affect players' perceived transparency of behavior moderation?

RQ1.2 Does punishment design affect players' perceived fairness of behavior moderation?

RQ1.3 Does punishment design affect players' coping strategies for punishments?

3.1.1 Moderation Transparency based on Punishment Design

Prior work has broadly recognized the importance of punishment notifications and explanations to improve moderation transparency (e.g., [25,99]). For example, researchers found that content moderation in localized communities like subreddits would become opaque when human moderators silently remove user content without notification or specifying reasons [52]. Especially many qualitative findings show that users perceive moderation as opaque when they do not receive notifications (e.g., [44,91]) and explanations (e.g., [67,73]) of punishments. Such perceived opacity will be intensified once punishment brings rippling effects to intervene in users' communication with online communities [27] and their online career development (e.g., income) [68]. Users thus want to obtain enough information about punishments. For example, on Facebook, they request explanations of why moderation systems issue inconsistent content removal decisions [98]. Especially a recent survey showed that custom messages specifying punishment reasons would increase users' perceived transparency of moderation [38]. In this sense, we assume that punishment notifications and explanations might allow players to perceive behavior moderation as more transparent than no notification or explanation:

H1.1: Punishment notification provision positively affects players' perceived transparency of behavior moderation, compared to no notification.

H1.2: Punishment explanation provision positively affects players' perceived transparency of behavior moderation, compared to no explanation.

Besides, punishment types may influence users' perceived transparency of moderation. Researchers have collectively understood that account suspension (i.e., permanent ban in game or de-platforming) is the most stringent punishment in moderation [46,56,67,73,91], where users lose the ability to continue using the original account to play games, post content, or communicate with others. Users might instantly generate perceived uncertainty and opacity toward moderation system due to such harsh punishment [73]. However, encountering relatively lighter punishments like content removal, users might take more time to make sense out of the punishments. They might develop perceived opacity more from the lack of notifications of punishment or limited direct communication with platforms than the punishment itself [44,67]. So, based on this line of prior work, we propose that compared to the relatively lighter punishments, such as content removal, other relatively heavier punishments, like permanent ban might negatively affect users' perceived transparency of behavior moderation:

H1.3: Experiencing (a) restricted access to game features (e.g., chat or matchmaking ban), (b) temporary ban, and (c) IP or permanent ban negatively affect players' perceived transparency of behavior moderation, compared to content or item removal.

3.1.2 Moderation Fairness based on Punishment Design

Platforms' transparency efforts, such as offering punishment notifications or explanations, have been seen as an important path to users' perceived fairness of moderation. For example, Ma and Kou found that content creators considered moderation as unfair when their videos were disproportionately hidden by YouTube's moderation algorithms without notifications [68]. Especially, creators felt the algorithms did not involve their voice in moderation decision-making procedures while moderation has already imposed negative effects on their channel performance and livelihoods [68]. Similarly, information disclosure of moderation might decrease such perceived unfairness. Jhaver et al. found that users receiving moderation explanations considered content removal as fair than no explanation on Reddit [45]. Vaccaro et al. uncovered that once explanations, either written by algorithms or humans, are offered, users' perceived fairness of account suspension on Facebook would increase [98]. This line of work shows how users will consider moderation fair if notifications and explanations are provided in moderation. Given the different dimensions of fairness as we discussed in Section 2.1, we propose that punishment notification and explanation provision can positively affect users' perceived fairness of behavior moderation:

H2.1: Punishment notification provision positively affects players' perceptions of (a) outcome fairness, (b) retributive justice, (c) procedural justice, and (d) restorative justice of behavior moderation, compared to no notification.

H2.2: Punishment explanation provision positively affects players' perceptions of (a) outcome fairness, (b) retributive justice, (c) procedural justice, and (d) restorative justice of behavior moderation, compared to no explanation.

While little prior literature directly points out the relationship between perceived fairness and behavior moderation, research in other fields predicts that punishments might negatively affect perceived fairness. For example, Xue et al. verified that in the enterprise context, users' perceived justice of punishments is negatively influenced by actual punishments [105]. Also, prior work has shown a positive correlation between transparency and fairness [63]. So, as we propose the negative relationship between punishments and perceived transparency, we predict the relationship between punishments and perceived fairness to be negative:

H2.3: Experiencing (1) restricted access to game features (e.g., chat or matchmaking ban), (2) temporary ban, and (3) IP or permanent ban negatively affect players' perceptions of (a) outcome fairness, (b) retributive justice, (c) procedural justice, and (d) restorative justice in platform governance, compared to content or item removal.

3.1.3 Coping Strategies based on Punishment Design

Prior moderation research has showed that more degree of transparency in moderation designs motivates more coping strategies adopted by moderated users. For example, when users do not receive moderation notifications, they might make cognitive efforts to conduct sense-making regarding why or how (e.g., algorithms or humans) punishments happen [67]. When users receive explanations that they perceive as unconvincing, they might request platforms to re-examine previous punishment decisions through appeal procedures [4,26,98]. Or they begin generating their own understanding and rationales to justify why they experience punishments [91]. They might make more behavioral efforts to contact platform representatives (e.g., human moderators) through third-party platforms if they fail to directly contact them [67,73]. Even for detachment factor, users who are informed of being blocked reacted to moderation with indifference and think moderation does not matter [49]. In game, when players are notified of permanent ban, some does not actively cope with it, while treating it as if nothing happened and buying a new account to commit toxicity [57]. Prior work has further showed that players can alter their behaviors

(e.g., toxic language) time after time, meaning that they might perform different behaviors in different moment of game [60]. Thus, we predict that the more actively platforms disclose information about moderation, the more diverse coping strategies moderated players will adopt, even though players might perform these strategies in different temporal patterns. We propose:

H3.1: Punishment notification provision positively affects players' adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking, compared to no notification.

H3.2: Punishment explanation provision positively affects players' adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking, compared to no explanation.

However, we predict punishment types, especially harsher ones, might not help users adopt diverse coping strategies. Much work has shown that severe punishments could work against player's positive behaviors (e.g., collaboration, reforming past behaviors). For example, a convicted person is not more likely to reform and improve their behaviors when they are punished by stronger punishments than weaker ones [10]. A recent large-scale experimental study with punished people also found that harsher punishments, such as prison sentences, were not more effective in helping convicts reform or preventing them from re-offending [42]. Similar situations happen in the context of content moderation. When experiencing relatively heavier punishments such as account suspension or community takedown, HCI researchers found users might not actively cope with punishments but become more toxic and hostile [43,94]. Game players who experience permanent account suspension also do not mean they become reformed players [57]. While under lighter punishments like content removal, users would generally decrease their frequency of posting spamming and hate speech [87,106]. Thus, we propose that compared to content or item removal, other relatively heavier punishments might negatively affect player's adoption of coping strategies.

H3.3: Experiencing (1) restricted access to game features, (2) temporary ban, and (3) IP or permanent ban negatively affect players' adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking, compared to content or item removal.

3.2 Perceived Transparency and Fairness as Predictors of Coping Strategies

As discussed, perceived transparency and fairness of moderation are two important user perceptions in the moderation literature. Prior work has also uncovered users' post-moderation efforts, such as appealing moderation decisions (e.g., [68,98]) or making cognitive efforts to make sense of why punishments happen [67]. However, we have relatively little knowledge of how punished users' perception of moderation is related to their behaviors afterward, especially in game. So, we ask:

RQ2 Do players' perceived transparency and fairness of behavior moderation affect their coping strategies for punishments?

Obtaining an initial understanding of this question, we found that prior work has recognized a positive relationship between perceived transparency and users' positive behaviors. For example, employees' perceived transparency of communication with employers positively affects employees' altruism and collaborations with others [50]. Users' perceived transparency of privacy policy is also a significant positive predictor of their cognitive trust in sharing health information with technologies [22]. In moderation context, several researchers have

indirectly uncovered that platform's transparency efforts (e.g., explanation provision) support users' positive behaviors. Jhaver et al. found that when content removal explanations were provided, users improved their behaviors, and thus fewer content removal cases happened to them [48]. Thus, we predict player's perceived transparency can support them in coping with moderation punishments:

H4.1: Players' perceived transparency of behavior moderation positively affects their adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking.

Similarly, we predict that perceived fairness can motivate players' adoption of coping strategies, which essentially are positive behaviors or cognitive efforts. That is because many prior studies have found that different dimensions of perceived fairness, including procedural justice, distributive justice, and interactional justice, positively affect people's organizational citizenship behaviors (e.g., [62,77]). Organizational citizenship behaviors represent a person's positive and constructive actions that can contribute optimally to organizations. So, perceived fairness plays a generally positive role in motivating people's efforts and positive attitudes. Especially, prior moderation literature has found a positive correlation between perceived fairness of moderation and productive user behaviors based on content removal explanation [45]. Increased perceived procedural justice has also been shown to decrease users' future behaviors of violating social media platform's content rules [96]. Reasonably, we propose a positive relationship between perceived fairness and coping strategies.

H4.2: Players' perceived outcome fairness of behavior moderation positively affects their adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking.

H4.3: Players' perceived retributive justice of behavior moderation positively affects their adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking.

H4.4: Players' perceived procedural justice of behavior moderation positively affects their adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking.

H4.5: Players' perceived restorative justice of behavior moderation positively affects their adoption of coping strategies for punishments, including (a) problem coping, (b) seeking social support, (c) detachment, (d) focusing on the positive, and (e) wishful thinking.

4 METHODS

The full survey is available as supplementary material, and we summarize our survey design, procedure, and sample in this section. Please note that we received 291 valid survey responses, but we removed one out of 291 from inferential statistical analysis because that participant's response did not meet the minimum size for inferential statistical analysis (see details in Section 5.2).

4.1 Survey Design

Our survey included three parts: (1) consent and screening, (2) punishment experience, and (3) demographics. In the consent and screening part, respondents read the consent sheet and indicated their agreement to participate. They were also asked about their age and experience with punishments (e.g., account, chat ban) from online

multiplayer games in the past. Participants who are under 18 years old or without experience with punishments were not given the option to proceed with the survey.

In the punishment experience part, respondents were first asked to multi-select the punishment types they had experienced with an option to manually type in more punishments. These punishment types were shared by many prior studies around social media moderation [73,98] and one recent work discussing punishments in game [66]. Once the participants made their multi-selection, a random punishment type from their multi-selection was presented. Participants were then asked to answer a set of follow-up questions on their perceived transparency and fairness on, as well as, coping strategies for the randomly presented punishment type that they had experienced. The reasons for the randomization were: (1) we wanted to focus only on the punishment that the participants had experienced so that their answers were not imaginary; (2) if the participants had experienced multiple punishments, it would be time-consuming to ask follow-up questions on each experienced punishment (i.e., # of questions * # of punishments) and the randomization could make the survey more efficient; (3) the randomization could control confounding factors that would bias the data. For example, the randomization could avoid participants reporting the punishments they remembered the most or the punishment they believed was the most unfair. In the last part of the survey, we asked about respondents' demographics, such as age, race, gender, and education levels. Also, we designed two attention check questions in different locations of our survey to ensure our data quality. Participants who failed these two attention check questions were excluded from our dataset.

4.1.1 Measurement Design

We measured respondents' perceived transparency, fairness, justice and adoption of coping strategies regarding the random experienced punishment. We reminded the respondents that the measurement questions were about the random experienced punishment by stating "please rate your agreement with the statements regarding the punishment decisions by [piped game]." Perceived transparency was measured by five items adapted from Gray and Durcikova's study [20] and Gonçalves et al.'s survey design [38]. The perceived fairness includes four dimensions, outcome fairness, retributive justice, procedural justice, and restorative justice, all of which were adapted from prior work. Outcome Fairness was primarily measured by three items from Colquitt's study [17] and Gonçalves et al.'s survey [38] and one item we made to summarize the outcome fairness notion. Retributive Justice was measured by five items adapted from Wenzel et al.'s study [102]. Procedural Justice was measured by five items adapted from Niehoff and Moorman's study [75], which measured perceptions of organization fairness. Restorative Justice was measured by five items adapted from Wenzel et al.'s study [102].

The factor group of coping strategies has five factors, including problem coping, detachment, social support, wishful thinking, and positivity. We adapted all survey items of these five factors from Scherer et al.'s study [83]. It is worth noting that originally, we adapted four items of wishful thinking from this work [83], but all these four items had significant cross loadings on the social support factor in EFA results. So, we removed this wishful thinking factor because all items did not converge into a single factor (see Section 4.4). Five-point Likert scales ranging from "strongly disagree" to "strongly agree" were used for all the items. The content of each item can be found in Table 2.

All the survey items were adapted to the context of gaming. For instance, the source of the punishment was changed to the particular game platform the respondent reported. For notions, including perceived transparency, outcome fairness, and coping strategies that are more related to personal and result-oriented notions, our survey items were adapted to follow the research trend that stresses subjective experiences of punishment (e.g., fairness

perceptions of account suspension punishment and reactions for it [44,47,91,98]). For example, for perceived transparency, one of our survey statements was "It is easy for me to see the status of punishments."

To reduce the social desirability bias, we used the method of proxy subjects to frame the survey statements that read negatively to respondents, namely the statements about justice notions, including retributive, procedural, and restorative justice. Social desirability is the tendency that study subjects, including survey respondents, tend to deny socially undesirable things which place the subjects in an unfavorable light [74,80]. One example is that a person could admit fewer violations of the law than actually committed. One method to deal with the social desirability bias is to use proxies, such as using someone who knows the respondents well [74] or similar others [34], instead of the target person. The social desirability bias might exist in the original statements of justice, i.e., "Overall, as a matter of justice, I should be punished", which might indicate negative and socially undesirable characteristics of the respondents. To mitigate the bias, we used "the convicted players" instead of "I" or "me" to avoid inquiring on whether punishment is desired or not by respondent personally and present respondents from mispresenting their punishment experience. This survey statement adaptation consideration was also supported by prior work that has assessed the perceived justice notions (e.g., [100,103]).

4.2 Procedure

After this study was approved by our institution's IRB office, we programmed our study design on Qualtrics. We first ran a pilot study with 15 respondents. These participants were compensated a $ 2 gift card (i.e., $12 hourly payment rate) for their participation. This pilot study helped us tweak some narratives of the questions and make the survey more readable and digestible to participants. Because of this change, after the pilot study, we did not include the data of the pilot study for the actual analysis. We then launched the survey on Prolific.co, an online participant recruitment service, to recruit players of online multiplayer games. The reason why we chose Prolific was that previous research has shown that Prolific offers high data quality for social science experiments and behavioral research [22,78]. To control the possible confounding factors (i.e., culture and country), we only recruited participants who (1) understood English, (2) experienced punishments in online multiplayer games, (3) and resided in the US. We compensated each respondent who finished the survey and passed attention check questions (N=291) with a payment rate of $12.54 per hour for completing the survey, which is higher than Prolific's site-wide average reward rate and the state minimum wage rate in the authors' state. The average time respondents used to complete the survey was around 14.8 minutes. The survey data collection was completed in May 2022.

4.3 Sample

We received a total of 432 responses, while 291 were complete and also passed our attention check questions. Please note that we randomized only one of the punishment types that respondents typed in for customizing the survey questions for them (see survey design in Section 4.1). We thus removed the response of one respondent out of 291 from further inferential statistical analysis because that respondent was the only one assigned to the punishment, "warning," which was only manually typed in by three respondents (see detail in Section 5.2). So here, we use a total of 291 valid responses for descriptive statistical purposes, while later, we will use 290 responses for referential statistical analysis. Table 1 summarizes the demographic information of the 291 respondents. Most players (65.64%) experienced restricted access to game features (e.g., chat ban, matchmaking restrictions) and temporary account ban. 52.92% of players experienced temporary ban, and 12.37% of players experienced

permanent or IP ban. Since many participants could report more than one type of punishment, the sum percentage of punishments experienced exceeds 100%, as shown in "Punishments Experienced" in Table 1.

Table 1. Player profiles (gender, education, age, race, and punishment types). All participants are from the US.

Gender	**Quantity (N=291)**	**Percent**
Female	103	35.40%
Male	173	59.45%
Non-binary / third gender	14	4.81%
Not Specified	1	0.34%
Education		
A high school diploma or equivalent	54	18.56%
Bachelor degree	94	32.30%
Doctoral degree	4	1.37%
Less than a high school diploma	8	2.75%
Master's degree	13	4.47%
Some college, no degree	73	25.09%
Two-year associate degree	45	15.46%
Race		
Asian	30	10.31%
Black or African American	18	6.19%
Hispanic, Latino, or Spanish	17	5.84%
Mixed race	33	11.34%
White or Caucasian	193	66.32%
Age	**Mean**	**Standard Deviation**
	31.26	0.55
Punishments Experienced	**Quantity**	**Percent**
Content or item removal	28	9.62%
IP or Permanent account ban	36	12.37%
Restricted access to game features	191	65.64%
Temporary account ban	154	52.92%
Warning	3	1.03%

4.4 Data Analysis

We performed exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and structural equation modeling (SEM) through Mplus, a statistical modeling program for researchers to analyze data. EFA was run firstly to check whether the factor structure we drew from prior work fit our survey data. In EFA, we used a robust weighted least-square estimator (WLSMV) and an oblique Geomin rotation method. The WLSMV estimator is better for ordered categorical indicators because it does not assume data in the factors to be normally distributed. After we got a valid factor structure from EFA, we further ran CFA to tone and build the final measurement model for the factors. We used the WLSMV estimator again in CFA and tested the convergent and discriminant validity of factors. Convergent validity will be supported if indicators (i.e., survey items) load significantly on the corresponding factor with standardized factor loading greater than 0.6, the Average Variance Extracted (AVE) higher than 0.5, and Cronbach's Alpha > 0.6. Discriminant validity will be supported if the correlation between factors is smaller than 0.85 and smaller than the square root of the AVE of each factor.

Based on the measurement model from CFA, we ran two SEMs with a WLSMV estimator to answer RQ1 and RQ2, respectively. SEM fits the measurement model and a set of linear regressions between factors. In the first SEM, we included all three independent variables, including punishment notification, explanation, and punishment types, as well as nine dependent variables, including a group of perceived fairness factors and coping strategy factors (we removed the "wishful thinking" factor after EFA, which was detailed in Section 5.1) and perceived transparency, as shown in Figure 1. The second SEM analysis involved five independent variables, including a group of perceived fairness factors and perceived transparency, as well as four coping strategy factors, as shown in Figure 2.

5 FINDINGS

This section will discuss how our findings answer RQ1 and RQ2. Answering RQ1, we found that punishment notification and explanations both significantly influenced players' perceived transparency of behavior moderation, and explanation provision significantly improved all notions of perceived fairness. However, nearly all facets of punishment design, including notification, explanation, and punishment types, did not affect how players cope with punishments with one exception. Explanation provision significantly affected players to adopt problem coping strategy. Answering RQ2, we found that both perceived transparency and fairness significantly affected players' adoption of coping strategies for punishments, while fairness notions like retributive, procedural, and restorative justice played more critical roles in affecting players to adopt more types of coping strategies than perceived transparency or outcome fairness. Additionally, in a casual inference logic, we found that explanation provision can only affect players to adopt problem coping strategy when they perceive behavior moderation as transparent.

5.1 Measurement Models

We ran Exploratory Factor Analysis (EFA) on all 10 factors that we adapted from prior work (see Section 4.3) to test whether our survey data would support this 10-factor structure. EFA results showed that a 9-factor structure had a better model fit than the original 10-factor solution. In the 10-factor solution, all four items of wishful thinking had significant cross-loadings on social support factor. Also, the factor loadings of these four items on wishful thinking were not at least two times higher than the loadings on the other factors. In other words, **the items of wishful thinking did not converge into a single factor**. Thus, based on EFA results, **we did not include wishful thinking for further analysis**. We correspondingly removed all hypotheses within H3 and H4 that involve wishful thinking (e.g., H4.1(e), H4.2(e)). The results of EFA further helped us remove a total of four items from three factors. In detail, the third item of procedural justice had significant cross-loadings on perceived transparency. Also, the first item of positivity, as well as the first and second items of detachment encountered, all had significant cross-loadings. Thus, we removed these four items to run CFA, as we applied strikethrough to them in Table 2.

Our construct had acceptant fit indices (RMSEA = 0.054, which is acceptable between 0.05 and 0.08 [24], 90% CI: [0.048, 0.058], CFI = 0.975 > 0.95, TLI = 0.972 > 0.95). Thus, the CFA results indicate that our construct has acceptable goodness of fit. The chi-square statistics were significant (χ^2= 1168.841, df = 629, p<.001). Usually, a chi-square test with a p-value greater than 0.05 (i.e., non-significance) shows a good model fit, and our results were contrary to what we expected. However, researchers (e.g., [9]) have broadly questioned the appropriateness of using chi-square test alone to evaluate the overall model file because it is sensitive to study's sample size and construct complexity. Thus, we alternatively used Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker–Lewis index (TLI) together [9] to describe the goodness-of-fit of our construct. Our construct had acceptant fit indices (RMSEA = 0.054, which is acceptable between 0.05 and 0.08 [24], 90% CI: [0.048, 0.058], CFI = 0.975 > 0.95, TLI = 0.972 > 0.95).

The convergent and discriminant validity of our measurement model is supported. First, nearly all factor loadings are greater than the requisite threshold of 0.6 [97] (see "factor loading" column in Table 2). One exception is the third item of detachment, which is smaller than 0.6. Therefore, we removed this item to run CFA again, and we strikethrough it in Table 2. We also reported each latent variable's Average Variance Extracted (AVE) and Composite Reliability (CR). Nearly all the AVEs are greater than 0.5, and CRs are greater than 0.6, which indicates good convergent validity. One exception is the AVE of detachment, which is 0.45 below the 0.5 threshold. However, the Composite Reliability (CR) of detachment is greater than 0.6, so the convergent validity of our

construct/measurement model is still adequate [30]. Besides, discriminant validity is supported for our measurement model. The correlations between factors are not only smaller than 0.85 but also smaller than the square root of AVEs, as shown in Table 3.

Table 3 Correlations between factors and the square root of AVEs. Note: each cell is the correlation coefficient between two factors with ∗ p < 0.05, ∗∗ p < 0.01, ∗∗∗ p < 0.001. All correlation coefficients are smaller than the corresponding square root of AVEs.

		1	2	3	4	5	6	7	8	9
1	Transparency									
2	Outcome Fairness	0.645***								
3	Retributive Justice	0.537***	0.608***							
4	Procedural Justice	0.626***	0.755***	0.657***						
5	Restorative Justice	0.23***	0.12 ns	0.3***	0.165***					
6	Problem Coping	0.395***	0.368***	0.479***	0.371***	0.389***				
7	Detachment	0.198***	0.171***	0.32***	0.325***	0.253***	0.239***			
8	Social Support	-0.163	-0.209	-0.102	-0.239	0.217***	0.078	-0.016		
9	Positivity	0.287***	0.394***	0.443***	0.471***	0.294***	0.598***	0.396***	0.108	
	The square root of AVE	0.867	0.951	0.847	0.858	0.804	0.824	0.670	0.849	0.879

Table 2 Factor loadings of the factors of punishment experience (CFA results). Note: since the survey questions were based on one random punishment participants experienced, [game] below represents where they experienced that punishment. Strikethrough refers to a survey item that has significant cross-loadings on other factors in EFA results.

	Factors (AVE CR)	**Survey Items**	**Factor Loadings**
Perceived Fairness	Transparency [20,38] (AVE=0.751 CR=0.937)	Overall, [game] tries to be transparent on punishment decisions.	0.921
		In general, I am notified about punishments from [game].	0.837
		It is easy for me to see the status of punishments.	0.726
		I am being told the reason behind punishments.	0.934
		Players like me are provided with information that is relevant to punishments.	0.899
	Outcome Fairness [17,38] (AVE=0.905 CR=0.974)	The punishments I got so far in [game] are fair so far.	0.962
		[game]'s punishment decisions are appropriate.	0.938
		The punishments I experienced are proportional to what I have done.	0.954
		[game] gave me the punishments I deserved.	0.951
	Retributive Justice [102] (AVE=0.717 CR=0.927)	Overall, as a matter of justice, the convicted players should be punished.	0.871
		Justice is served at the moment that the convicted players are punished in [game].	0.844
		The only way to restore justice is to punish the convicted players.	0.86
		The convicted players deserve to be penalized.	0.876
		For the sake of justice, some degree of suffering has to be inflicted on the convicted players.	0.78
	Procedural Justice [75] (AVE=0.736 CR=0.917)	The punishment decisions are made by [game] in an unbiased manner.	0.811
		To make the punishment decisions, [game] collects accurate and complete information.	0.903
		~~[Game] clarifies punishment decisions and provides additional information when requested by players~~	~~0.613~~
		All punishment decisions are applied consistently across all affected players.	0.804
		The decision-making process of punishments has followed ethical and moral standards.	0.908
	Restorative Justice [102] (AVE=0.646 CR=0.916)	For justice to be reinstated, [game] needs to achieve agreement about the values violated by the players.	0.87
		To restore justice, the players and [game] need to reaffirm consensus on the values and rules.	0.833
		Without the players' sincere acknowledgment of having acted inappropriately, the injustice is not completely restored.	0.669

		A sense of justice requires that the players and [game] develop a shared understanding of the harm done by players' behaviors.	0.785
		Justice is restored as soon as the player has learned to endorse the values violated by their behaviors.	0.785
		For a sense of justice, players and [game] need to reaffirm the belief in shared values.	0.864
Coping strategies	Problem Coping [83] (AVE=0.679 CR=0.913)	I know how to improve my behavior to avoid punishments in the future.	0.833
		I try to analyze punishments in order to understand them better.	0.748
		I could make a plan of action and follow it to improve my behaviors.	0.902
		I could come up with a couple of different solutions to punishments.	0.754
		I could analyze the punishments in order to understand them better.	0.871
	Detachment [83] (AVE=0.449 CR=0.619)	~~After punishments, I usually continue playing [game] as if nothing happened.~~	~~0.445~~
		~~I was just unlucky to be punished by [game].~~	~~0.536~~
		~~I try to forget about the punishment I received.~~	~~0.158~~
		I feel that time will make a difference; the only thing to do is wait	0.652
		I usually wait to see what will happen from punishments before I do anything	0.687
	Social Support [83] (AVE=0.721 CR =0.911)	I tend to talk to someone about the punishments I experience in [game].	0.8
		I tend to ask someone I trust for advice about the punishments.	0.973
		I tend to talk to someone who could do something concrete for my punishments.	0.843
		I want to receive sympathy and understanding from someone.	0.765
	Positivity [83] (AVE=0.772 CR=0.910)	~~Overall, I tend to focus on the positive after punishments.~~	~~0.703~~
		The punishment can actually inspire me to do something positive.	0.816
		The punishment encouraged me to discover what is important in playing [game].	0.918
		The punishment causes me to grow or change in a good way.	0.899

5.2 Punishment Design

Given the variety of punishments each participant experienced, we randomly assigned one of the multi-selected punishments to further probe the context of the punishment. The most frequent randomly assigned punishments were restricted access to game features and temporary account ban, as shown in Table 4. Warning, as a type of punishment, was assigned to only one participant who chose to freely specify the punishment experienced, which was "warning." Thus, to make sure each punishment has enough data points for the Structural Equation Modeling (SEM) analysis, we removed this individual response. Our final dataset then contained a total of 290 valid responses for further analysis.

Every participant answered the game where they experienced the randomly assigned punishment (see "Game Platform" column in Table 4). Many participants mentioned the game platforms beyond the options we provided. For example, they experienced restricted access to game features in games of Ancient Anguish, Homeworld, Lineage 2, and more. Also, others reported their temporary account ban happened in Alliance of Heroes, Gears of war, Magic: The Gathering Online, Ragnarok Online, Sea of Thieves, and more. After they answered the game platform questions, the rest of the questions were automatically customized by the game name they selected/typed.

Table 4. Punishment design: The randomly assigned punishment and game where participants experienced it.

Randomly Assigned Punishments (Quantity, Percent)	**Game Platforms**	**Quantity**
Warning (1, 0.34%)	Roblox	1
Content or item removal (11, 3.78%)	World of Warcraft (WoW)	3
	Apex Legends	1
	Final Fantasy	1
	Fortnite	1
	Grand Theft Auto	1
	League of Legends	1
	Minecraft	1
	Roblox	1
	Runescape	1
Restricted access to game features (138, 47.42%)	League of Legends	18
	Other game	16
	Fortnite	11
	Apex Legends	10
	World of Warcraft (WoW)	9
	Minecraft	8

	Overwatch	8
	Dead By Daylight	7
	Rocket League	7
	Dota 2	6
	Call of Duty	5
	Runescape	5
	Halo	4
	Final Fantasy	3
	Rainbow Six Siege	3
	Valorant	3
	Counter Strike Global Offensive (CS:GO)	2
	Destiny 2	2
	Smite	2
	Splatoon 2	2
	World of Tanks	2
	Battlefield	1
	Club Penguin	1
	New World	1
	Pokémon go	1
	Team fortress 2	1
IP ban or Permanent account ban (20, 6.78%)	Other game	4
	Minecraft	2
	PUBG	2
	Apex Legends	1
	Battlefield	1
	Call of Duty	1
	Club Penguin	1
	Counter Strike Global Offensive (CS:GO)	1
	Fortnite	1
	Gaia Online	1
	League of Legends	1
	NBA 2k	1
	Runescape	1
	Team fortress 2	1
	World of Warcraft (WoW)	1
Temporary account ban (122, 41.92%)	Other game	27
	League of Legends	21
	World of Warcraft (WoW)	21
	Fortnite	11
	Minecraft	8
	Call of Duty	6
	Apex Legends	4
	Counter Strike Global Offensive (CS:GO)	3
	Halo	3
	Dota 2	2
	Final Fantasy	2
	Grand Theft Auto	2
	Overwatch	2
	Runescape	2
	Club Penguin	1
	Gaia Online	1
	NBA 2K	1
	Pokémon go	1
	PUBG	1
	Rainbow Six Siege	1
	Roblox	1

As Table 5 shows, many players reported they received punishment notifications and explanations. 248 out of 291 participants reported that they were notified of the punishment by the games. These participants received notifications by emails and pop-up windows or in-game messages. Also, 217 out of 291 participants reported that they received punishment explanations. However, 74 out of 291 participants reported they did not receive or were not sure if receiving explanations. 26 out of 291 participants reported that game platforms did not explain why punishment happened to them even though receiving notifications. Furthermore, seven out of 291 participants reported that they did not receive punishment notifications but received explanations. Their responses to open-ended questions indicated that many of them found explanations when logging into game, while the game did not notify them of punishment beforehand. For example, they said: "When I tried to log in one day, it told me that I was temporarily suspended from logging in to my account." Another similar response was, "Could not login or play upon trying." These initial qualitative findings prompted us to dive deeper to understand how punishment design like

notification and explanation would affect the ways how players perceive behavior moderation and moderation punishments.

Table 5. Punishment design: Self-reported punishment notification and explanation provision.

		Explanation Provision			
		Yes	No	Not sure	Total
Notification Provision	Yes	204	26	18	248
	No	7	20	2	29
	Not sure	6	2	6	14
	Total	217	48	26	291

5.3 RQ1: Punishment Design → Transparency/Fairness Perceptions & Coping Strategies

To answer RQ1 and its subsequent RQ1.1 to RQ1.3, we built an SEM to model the hypothesized relationships between different types of punishment designs (punishment types, notification, and explanation) and (1) perceived transparency, (2) perceived fairness, and (3) intent to adopt coping strategies. This SEM model has a good model fit: $\chi 2$= 1331.714, df = 789, p<.001; RMSEA = 0.049 < 0.05 [24], 90% CI: [0.044, 0.053], CFI = 0.969 > 0.95, TLI = 0.964 > 0.95.

Table 6 summarizes the SEM results, as well as whether each hypothesis is fully or partially supported (see "Results" column). Overall, punishment types do not have a significant association with players' perceived transparency and fairness or their coping strategies. Providing punishment notification and explanation generally has positive effects on players' perceived transparency and fairness. We will elaborate on these results in the next subsections.

Table 6. The first SEM (hypothesis testing) results for RQ1 (model 1). Note: The solid arrows (→) present significant relationships, and broken arrows (⇢) represent tested relationships that are non-significant. (+) or (-) indicates a positive or negative effect between factors. Coefficient β with ∗ p < 0.05, ∗∗ p < 0.01, ∗∗∗ p < 0.001. n.s. means non-significant.

Model 1					
RQ	**Hypothesis#**			**Coef (β)**	**Results**
RQ1.1	H1.1		Notification provided → Perceived Transparency (+)	0.543***	Fully support
	H1.2		Explanation provided → Perceived Transparency (+)	1.537***	Fully support
	H1.3	*baseline*	*Content or item removal*		No support
		(a)	Restricted access to game features ⇢ Perceived Transparency (+)	0.268 n.s.	
		(b)	Temporary account ban ⇢ Perceived Transparency (+)	0.034 n.s.	
		(c)	IP or permanent ban ⇢ Perceived Transparency (-)	-0.094 n.s.	
RQ1.2	H2.1	(a-d)	Notification provided ⇢ Outcome Fairness (+), Retributive Justice (+), Procedural Justice (+), Restorative Justice (-)	n.s.	No support
	H2.2	(a)	Explanation provided → Outcome Fairness (+)	0.773***	Fully support
		(b)	Explanation provided → Retributive Justice (+)	0.475*	
		(c)	Explanation provided → Procedural Justice (+)	0.645**	
		(d)	Explanation provided → Restorative Justice (+)	0.358*	
	H2.3	(1&a-d)	Restricted access to game features ⇢ Outcome Fairness (+), Retributive Justice (+), Procedural Justice (+), Restorative Justice (+)	n.s.	No support
		(2&a-d)	Temporary account ban ⇢ Outcome Fairness (-), Retributive Justice (-), Procedural Justice (+), Restorative Justice (+)	n.s.	
		(3&a-d)	IP or permanent ban ⇢ Outcome Fairness (-), Retributive Justice (-), Procedural Justice (+), Restorative Justice (+)	n.s.	
RQ1.3	H3.1	(a-d)	Notification was provided ⇢ Problem Coping (-), Social Support (+), Detachment (-), Positivity (-)	n.s.	No support
	H3.2	(a)	Explanation was provided → Problem Coping (+)	0.55**	Partially support
		(b-d)	Explanation was provided ⇢ Social Support (-), Detachment (+), Positivity (+)	n.s.	
	H3.3	(1&a-d)	Restricted access to game features ⇢ Problem Coping (+), Social Support (-), Detachment (+), Positivity (+)	n.s.	No support

(2&a-d)	Temporary account ban ⇢ Problem Coping (+), Social Support (-), Detachment (+), Positivity (+)	n.s.
(3&a-d)	IP or permanent ban ⇢ Problem Coping (+), Social Support (-), Detachment (+), Positivity (+)	n.s.

5.3.1 RQ1.1: Punishment Design → Perceived Transparency of Behavior Moderation

The effects of notification and explanation provision on participants' perceived transparency of behavior moderation are significantly positive. **This supports H1.1 and H1.2**. When participants are provided with punishment notifications, they consider behavior moderation as more transparent (β = 0.534***). When participants are provided with punishment explanations, they consider behavior moderation as more transparent (β = 1.537***). Especially, the size of the coefficient of punishment explanation provision is greater than the one of notification provision, indicating punishment explanation plays a greater role in improving the perceived transparency than notification. Besides, punishment types do not significantly affect perceived transparency. Thus, **H1.3a-c are not supported**.

5.3.2 RQ1.2: Punishment Design → Perceived Fairness of Behavior Moderation

The effects of explanation provision on outcome fairness (H2.2a), retributive justice (H2.2b), procedural justice (H2.2c), and restorative justice (H2.2d) are all significantly positive. Besides, neither notification provision nor punishment types have a significant effect on any dimension of perceived fairness. **H2.3(1-3&a-d) and H2.1a-d are thus not supported**. In sum, punishment explanations play an important role in affecting whether players perceive behavior moderation as fair, while notification and punishment types do not affect, which answers RQ1.2.

5.3.3 RQ1.3: Punishment Design → Coping Strategies for Punishments

Nearly no punishment design has significant effects on players' coping strategies for punishments, with one exception. When games provide explanations, players are more likely to initiate "problem coping" to address the punishments (β = 0.55**), **supporting H3.2a**. Other than this, transparency efforts from games, such as providing punishment notifications or explanations, generally did not motivate players to cope with the punishments by adopting coping strategies such as seeking social support, focusing on the positive, or detachment. Besides, punishment types are not key factors for players to decide on coping strategies. **H3.1a-d**, **H3.3(1-3&a-d), and H3.2b-d are not supported**.

Taken together, to answer RQ1, we found that when either punishment notifications or explanations are offered, players would consider behavior moderation as transparent. Especially, explanation provision can significantly improve players' different dimensions of perceived fairness such as restorative and procedural justice as well as affect players to adopt problem coping actions for punishments. However, punishment types as one of the punishment designs did not have significant effects on either players' perceived transparency, fairness, or coping strategies.

5.4 RQ2: Perceived Fairness and Transparency → Coping Strategies

Our second SEM (as shown in Table 7) has a good model fit: $\chi 2 = 1120.819$, df = 593, $p<.001$; RMSEA = 0.055, which is acceptable between 0.05 and 0.08 [24], 90% CI: [0.05, 0.06], CFI = 0.975 > 0.95, TLI = 0.972 > 0.95. This model tests whether and how players' perceived transparency and fairness of behavior moderation affect players' coping strategies for punishments.

Table 7. The second SEM (hypothesis testing) results for RQ2 (model 2). Note: The solid arrows (→) present significant relationships, and broken arrows (⇢) represent tested relationships that are non-significant. (+) or (-) indicates a positive or negative effect between factors. Coefficient β with ∗ p < 0.05, ∗∗ p < 0.01, ∗∗∗ p < 0.001. n.s. means non-significant.

Model **2**					
RQ	**Hypothesis#**			**Coef β**	**Results**
RQ2	H4.1	(a)	Perceived Transparency → Problem coping (+)	0.168*	Partially support
		(b-d)	Perceived Transparency → Social Support (-), Detachment (+), Positivity (-)	n.s.	
	H4.2	(a-d)	Outcome Fairness ⇢ Problem coping (+), Social Support (-), Detachment (-), Positivity (+)	n.s.	No support
	H4.3	(a)	Retributive Justice → Problem coping (+)	0.303***	Partially support
		(d)	Retributive Justice → Positivity (+)	0.176*	
		(b,c)	Retributive Justice ⇢ Social Support (+), Detachment (+)	n.s.	
	H4.4	(a)	Procedural Justice ⇢ Problem Coping (-)	-0.015 n.s.	Partially support
		(b)	Procedural Justice → Social Support (-)	-0.223*	
		(c)	Procedural Justice → Detachment (+)	0.337**	
		(d)	Procedural Justice → Positivity (+)	0.32**	
	H4.5	(a)	Restorative Justice → Problem Coping (+)	0.261***	Fully support
		(b)	Restorative Justice → Social Support (+)	0.268***	
		(c)	Restorative Justice → Detachment (+)	0.169*	
		(d)	Restorative Justice → Positivity (+)	0.204***	

We answer RQ2 by uncovering that perceived transparency positively affects problem coping (β = 0.168*, H4.1a) but does not affect other coping strategies (H4.1b-d). So, **H4.1 is partially supported**. While perceived fairness generally has positive influences on coping strategies, perceived outcome fairness (H4.2a-d) does not have significant effects on players' coping strategies. **H4.2 thus is not supported**. Other than perceived outcome fairness, perceived retributive justice positively affects problem coping (β = 0.303***, H4.3a) and positivity (β = 0.176*, H4.3d) but does not significantly affect social support (H4.3b) and detachment (H4.3c). **H4.3 are thus partially supported**. Furthermore, while perceived procedural justice has a positive effect on detachment (β = 0.337, H4.4c) and positivity (β = 0.32**, H4.4d), it has a negative effect on social support (β = -0.223*, H4.4b), which is contrary to what we hypothesized, and does not affect problem coping (H4.4a). Thus, **H4.4 is partially supported**. Then, perceived restorative justice has a positive effect on all types of coping strategies, including problem coping (β = 0.261***, H4.5a), social support (β = 0.268***, H4.5b), detachment (β = 0.169*, H4.5c), and positivity (β = 0.204***, H4.5d). Thus, **H4.4 is fully supported**.

In sum, answering RQ2, we found that when players perceived behavior moderation as transparent, they tended to proactively cope with punishments (i.e., problem coping strategy). When players think behavior moderation is conducted in an unbiased manner (i.e., procedural justice), they are more likely to adopt detachment and positivity to cope with punishments but less likely to seek social support. And when players think behavior moderation is conducted in a correct, punitive manner (i.e., retributive justice), they tended to adopt problem coping and positivity to cope with punishments. Last, when players think the game platform affirms consensus on its values and rules with them (i.e., restorative justice), they would be more likely to adopt all coping strategies for punishments.

5.5 Perceived Fairness and Transparency Have No Mediation Effects with One Exception

Since punishment design affects perceived transparency and fairness and also perceived transparency and fairness affect coping strategies, it is possible that perceived transparency and fairness serve as mediators between punishment design and coping strategies. To test this, we ran three more SEMs to test if introducing mediators (i.e., perceived transparency and fairness) would make the significant effects of punishment design on coping strategies

insignificant [41]. Since only the effect of explanation on problem coping was significant (see RQ1.3 in Table 6), we thus tested the mediation effect on this path only by introducing (1) perceived transparency and fairness both as mediators, (2) perceived transparency as the only mediator, and (3) perceived fairness as the only mediator between punishment design and problem coping. Since other effects of punishment design on coping strategies are insignificant, there is no point in testing mediation for these paths.

Model (1) and (3) have poor model fits: RMSEA > 0.1, CFI < 0.9, and TLI < 0.9, thus, do not support a mediation model. However, Model (2) has a good model fit: RMSEA > 0.054, which is acceptable between 0.05 and 0.08 [24], 90% CI: [0.046, 0.062], CFI = 0.972 > 0.95, and TLI = 0.964 > 0.95. When perceived transparency is introduced as a mediator between punishment design and coping strategies, the direct effects of punishment designs on coping strategies, including the only significant one (explanation → problem coping), are no longer significant, indicating that **perceived transparency fully mediates the relationship between explanation and problem coping** (shown in Figure 3). If players are given punishment explanations, they will perceive more transparency of the behavior moderation and inherently are more likely to improve the adoption of problem coping strategy for punishments.

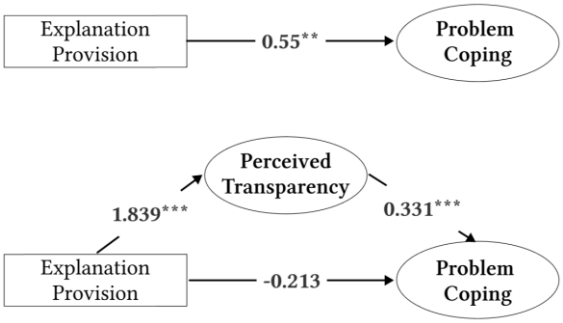

Figure 3 Perceived transparency fully mediates the relationship between explanation and problem coping with ∗ p < 0.05, ∗∗ p < 0.01, ∗∗∗ p < 0.001.

6 DISCUSSION

We conducted a survey study to understand how players experience the punishment design of behavior moderation in multiplayer online games, identifying interrelationships among three facets of punishment design, players' perceived fairness and transparency of moderation decisions and players' intended adoptions of coping strategies for punishments. In this section, we will discuss how our findings help deepen understanding of behavior moderation in the context of online games. Then, we will discuss how we should consider moderation experience as part of player experience and derive practical implications for moderation design and policymaking from our findings.

6.1 Extending Understanding of Moderation Experience in the Context of Online Gaming

Prior work has focused broadly on understanding users' experiences with moderation systems on social media such as Reddit [45,48], YouTube [67,68], Facebook [98], and more. As our study showed, game players are also concerned about the issues such as transparency or fairness of moderation systems that many HCI researchers have discussed in the social media context (e.g., [47,51,99]).

First, our findings helped quantitatively confirm the importance of punishment notifications and explanations to improve the perceived transparency and fairness of moderation in the context of online gaming. Prior work has found that social media users perceive the opacity of moderation decisions on social media because they do not receive notifications or explanations from platforms (e.g., [26,67,73,91]). When users receive moderation explanations such as appeal explanations [98] or reasons for account suspension [45], their perceived transparency and fairness would be improved. Resonating with this line of work, we found the positive effects of punishment notification and explanation on players' perceived transparency. Importantly, extending the prior work, we found explanation provision played a more critical role than notification in affecting both perceived transparency with a larger effect size and all notions of perceived fairness. That said, when online multiplayer games construct more transparent, fairer moderation systems, specifying why players experience punishments would be key to helping they understand (1) punishment as fair (i.e., outcome fairness), (2) the punitive logic of moderation system as legitimate (i.e., retributive justice), (3) the procedures of moderation decision-making as justified (i.e., procedural justice), and (4) games confirm the rules and values within the same page with them (i.e., restorative justice).

Second, only punishment explanation from the three punishment design components directly motivates players to actively cope with punishments and indirectly drives players to do so through perceived transparency. Prior work has uncovered that users might conduct behavioral or cognitive efforts to avoid or resist punishments on social media (e.g., [4,26,91,98]) but did not explicate why users initiate such coping efforts for punishments. Our study specifies one motivation: When players receive punishment explanations, they would be more likely to analyze the punishments, improve past behaviors, or make a plan to handle both (i.e., problem coping strategy). So, by moving beyond prior work that stresses the importance of moderation explanations [44,67,98], we emphasize that explanation design is important not only because of its impacts on improving perceived transparency of moderation but also driving punished users to actively handle the negative effects of punishments.

Third, punishment types as one of the punishment design components do not play a role in affecting the perceived transparency and fairness as well as intended adoption of coping strategies, which are somewhat different from what we expected. We conjecture that online gaming culture is different from social media platforms in its closedness from scrutiny of the outside world and corporate owners enjoying much power in making authoritative decisions that are rarely challenged, and players are accustomed to this culture and rarely challenge the severity of punishment [55].

While some HCI research showed that severe moderation decisions such as account suspension impact social media users' moderation experiences (e.g., fairness perceptions [44,98]), our results showed that players paid less attention to punishment severity compared to punishment designs (e.g., notification, explanation). Such new understanding of moderation experience exactly showed that in the context of competitive online multiplayer games, where toxic behaviors are prevalent and can be influenced by game designs (e.g., players' powerless in matchmaking [56]), players might have already normalized toxicity and become insensitive to punishment severity [5,57]. This suggests the importance of punishment design – a design that could inform players of what procedure the decision-maker, i.e., game platforms, conducts to make punishment decisions and sequentially how to help players reform behaviors [58] if they truly violate platform policies. So, not like social media users who request explanations and notifications for certain punishments (e.g., account suspension [26,44,98], revenue deduction [67]), game players purely request explanations to understand how punishment decisions are made, which directly speaks to the perceived justice of moderation they desired.

6.2 Foregrounding Justice Notions in Investigating Moderation Experiences

Although this work is focused on the context of online gaming, our findings about perceptions and experiences of justice can form meaningful conversations with moderation research in other contexts, and thus deepen our general understanding of moderation. First, many HCI researchers have drawn from the notion of procedural justice to call for moderation system to increase people's perceived fairness of moderation decisions [85] and increase their participation in the moderation decision-making process [26], as well as to issue consistent moderation decisions across time, users, and content policies [68]. Connecting with this body of work, which is focused on social media contexts, we offer a new understanding of users' actions after they perceived the procedural justice of moderation in game: Players were more likely to adopt detachment and positivity but less likely to adopt social support. That means, punished players treated punishment more as a typical procedure they would go through rather than an emotionally challenging event. In contrast, when players perceived little procedural justice in moderation decision-making (i.e., game platforms offer limited resources to help players understand the legitimacy of punishment decision-making), they were more likely to seek help from peers, i.e., social circles. This finding resonates with prior work that when punished users on social media consider their voice is not involved in moderation decision-making, they will ask for community support to make sense out of or learn about punishments [26,67]. And importantly, our findings convey an important message: designing moderation procedures that involve punished players' voice and participation can not only enhance perceived procedural justice of moderation but also decrease the chances of the perceived moderation unfairness being generated and disseminated by players.

Moreover, as game and social media platforms typically adopt a retributive/punitive justice logic on convicted users through punishments [36,56,73], our findings confirmed its effectiveness. When players perceived the retributive justice of moderation, believing that penalties were fairly issued, they would adopt problem coping and positivity, two out of four coping strategies for punishments, indicating the fairly good effects of perceived retributive justice on helping players reflect or reform behaviors. Also, players upholding this notion would adopt more problem coping than focusing on the positive, meaning that players accept penalty as a problem that they should address instead of ignoring it. Thus, the punitive logic still works in influencing players so that they take punishments seriously and seek to reform. And such effectiveness of retributive justice further iterates the benefit of taking procedural justice into account in punishment design [26,54], as our findings showed a high positive correlation between procedural and retributive justice, compared to other pairs in Table 3.

Beyond designing for both procedural and retributive justice, our findings confirmed the importance of restorative justice in punishment design, which has been advocated by studies of non-gaming contexts (e.g., [85,98,104]). We found that when players perceived restorative justice, such as efforts and resources that help them reform, they were likely to adopt all types of coping strategies. And importantly, if games only ensure the perceived retributive and procedural justice but not restorative justice, players will still look for social support to cope with punishments. That means, sometimes, social support is important for punished users to better understand moderation decisions [26,67] but could occasionally lead to collective circumventing moderation decisions or gaming moderation systems [14,35]. Thus, online games need to ensure perceived fairness, including perceived retributive, procedural, and restorative justice, to effectively help players reflect and cope with punishments.

Punished players' diverse needs for justice, when taken into consideration together with prior findings on punished users' needs for justice in other game-related and non-gaming contexts (e.g., [44,69,99]), raise a critical question regarding general moderation research and practice – how moderation design could conceive punished users as an important stakeholder group. From platform's perspective, punished players are deemed to be

offenders who violate platform policies (e.g., code of conduct). And researchers would leverage this perspective to design moderation justice that values offenders' participation and voice (e.g., [85,104]). While from punished players' perspectives, the power imbalance between game platforms and players in punishment decision-making is apparent, where players experience punishment and bear with its negative impacts on their player experience. Especially, as our findings showed that games did not explain well what and why players were accused of, these punished players would be socially stigmatized with a label or stereotype of toxic players or offenders [57]. However, like many other social media users, players might encounter hardships of contesting punishment decisions [69,98,99] and justifying the punishments are legitimate on their own force [79]. Thus, users are less motivated to put effort into clearing their name, if they perceive the punishment decision-making as lacking in justice.

Although sometimes, punished users can find social support to make sense of punishments, this is still extra labor and could be attributed to inadequate and ineffective punishment design. Like users' behaviors in audio-based communities [51], players' behaviors might be complex and nuanced, that voluntary human moderators find tricky to adjudicate. Game publishers could do more to educate and instruct punished players, as more researchers have called for platform moderation to take more responsibilities such as incorporating education in moderation (e.g., [47,73]). Our findings pointed out a pragmatic way – designing better punishment explanations. That is because, as we found, without sufficient or informative explanations, players would not consider behavior moderation as fair in terms of all justice notions, including procedural, retributive, and restorative justice. But currently, relatively little work has started to design moderation explanations except for several situated in the social media context [47,98], so we call for more HCI researchers' attention on explanation design for more transparent and fairer moderation in broader contexts.

6.3 Moderation Experience as Part of Player Experience

Player experience (PX) research has growing attention to toxicity in online games, as well as moderation techniques that could curb toxicity [2,5,55,89]. Moderation experience is the other side of the coin, concerning how those moderation techniques impact players who are considered as toxic. In this regard, moderation experience unambiguously belongs to player experience. When players engage in online games, they interact with numerous, interlocking systems, among which some govern the core gameplay, some manage interpersonal communication, coordination, and teaming, and some control behavior moderation systems. Although not part of the core gameplay of a game, moderation experience cuts across many facets of PX, such as social experiences (e.g., a player is temporarily losing the ability to communicate with a chat restriction or seeks social support from their fellow players), emotional experiences (e.g., a player is frustrated due to not understanding an account suspension), and player engagement (e.g., a player is no longer able to play if their account is suspended).

Importantly, the purpose of our study is not to refute the necessity and legitimacy of behavior moderation in multiplayer online games. Rather, we are to identify punished players as a unique player group that needs more scholarly and design attention. Indeed, we see many connections between players' interactions with punishments and PX. When put in an adverse situation (i.e., being punished), players have emergent needs. The self-determination theory (SDT), widely used in HCI game research [95], establishes three core needs as autonomy, competence, and relatedness. The SDT holds relevance for us to understand moderation experience in our study. First, the punished player has the 'autonomy' need as they want to make decisions on their own, actively addressing the problem of moderation penalty. Certainly, punishment design affects this autonomy need. For example, the

design of moderation explanation could enhance players' autonomy and encourage them to take the problem-solving route; and better procedural justice in punishment design could facilitate certain aspects of autonomy while inhibit others. Second, the punished player has the 'competence' need, clearly shown in our findings about players' desire for more punishment information. Information helps grow their competence in areas such as knowledge about moderation decision-making processes, as well as normative standards for in-game behavior. Third, the punished player has the 'relatedness' need, manifest in the social support coping strategy where people are punished and subsequently turn to others for social support. However, such relatedness need could be less if there is sufficient procedural justice so that players could count on the system to make the right decisions.

6.4 Implications for Design: Rethinking Moderation Design in Multiplayer Online Games

Multiplayer online games usually follow a rudimentary, punitive model in player behavior moderation, issuing a penalty and expecting the punished player to either reform or leave the game. As a pertinent example, the Fair Play Alliance [76], a global coalition of game companies working together to promote healthy and safe gaming environments, frames their primary solution to toxicity in languages such as planning and building "a penalty & reporting system" in their recent Disruption and Harms in Online Gaming Framework [25]. The punitive model has severe limitations in such dimensions as transparency and fairness [47,98], as demonstrated by moderation researchers (most often in the context of social media moderation). Bridging the moderation literature and the HCI game literature, our work points to the importance of moderation design, especially in terms of providing explanations and notifications. Without sufficient information, it could be challenging for players to understand why they are punished or to act accordingly. As a result, simply sending a moderation penalty fails to realize the full potential of creating a teachable moment [72] for players who have committed toxicity, rendering a poor player experience.

More problematic is the situation when the moderation decision is unjust, but the punished player has nowhere to resort to. Since justice perceptions such as fairness and transparency affect players' coping strategies, it is reasonable to assume that perceived injustices in moderation decisions reversely affect players' coping actions. In other words, moderation design's insufficient information provision lowers players' fairness and transparency perceptions, which in turn could reduce their willingness and action to improve their future in-game behaviors. Our findings provided empirical support for this observation: when perceived transparency and fairness are low, players would count on their fellow players for help, but better transparency and fairness could enable players to seriously consider their penalties and take actions to reform their future behavior.

Moving beyond the simplistic moderation design, we could rethink punishment design by drawing inspirations from game design. Video games are known for presenting players with a challenge in game and then supporting players to overcome it [53]: To defeat the final boss, the players are well prepared through the accumulation of experience points, equipment, and the improvement in game mechanics and knowledge (i.e., the needs of autonomy and competence). In multiplayer online games, players team up with others to accomplish a larger goal (i.e., the need for relatedness). But if we consider punishment as a challenge, then players are left on their own to cope with the challenge. Clearly, there is a large gap where player needs could be met if moderation design utilizes what we have already learned in PX about helping players to overcome challenges. By drawing this analogy, we suggest that punishment could be productively reframed as a challenge and call for better design that could help players to overcome this challenge.

Specifically, our findings highlight several dimensions to rethink moderation design in multiplayer online games: First, the *informational* dimension deals with what information should be provided alongside a punishment and in what way. Our findings showed that if players do not understand why they are punished, they could struggle to improve. In cases where they are wrongly convicted, access to the rationale behind their punishment is even more important. Our findings suggested that informational provision significantly impacts players' experiences with moderation decisions. Explanation provisions could become a teachable moment and trigger players' subsequent actions of problem-solving. Thus, behavior moderation systems could consider providing explanations when issuing a penalty. Specifically, explanation design should also consider the level of granularity and detailedness. When an explanation only refers to a vague community guideline, players would still struggle to analyze their deeds [58]. Explanation design could include both high-level pointers to specific policies violated as well as precise mappings between the policies and specific player behaviors in question. It could also be helpful if players are encouraged to discuss their penalties with fellow community members through collective sensemaking.

Second, the *social* dimension deals with players' relatedness needs. Our findings found several occasions where players would turn to social support as a coping strategy. When players are punished, they could be empowered to connect with fellow players to better figure out how to overcome the negative event.

Third, the *temporal* dimension takes a developmental view of players' moderation experience. While most moderation design stops at issuing a penalty, it is where punished players start to experience, feel, and react. These experiences are currently unaccounted for in the moderation design and thus a missed opportunity. Thus, restorative means could be designed around existing moderation systems. For example, various forms of player support could be designed for punished players. More mechanisms could be built where punished players could be connected to helpful resources that help them learn behavioral standards and other community members who are willing to offer social support.

7 LIMITATIONS AND FUTURE WORK

Our study does not aim to define what punishment design components, including punishment explanation and notification as well as punishment types, look like in real games, especially as our survey respondents reported many online games. Even though we used explicit language like "notification" and "explanation (i.e., reasons)," we did acknowledge that players situated in different online games might conceptualize punishment design differently. Thus, future work could explore and co-design with punished players to understand what consists of a moderation explanation or notification that can better center around players' best interests. Also, even though our sample size fit the minimum standard (e.g., n>200) for running factor analysis and SEM [33], we did recognize possibilities for further work to study with more players.

We did not aim to assess whether and how players' perceptions and actions would be different among different game genres or types, because existing literature has recognized that there lacks a consensus on a taxonomy of game genres, and game categorization methods might contain certain subjectivity or lack clarity [16,43]. For example, people might consider Overwatch either a shooter or a fight game. Or it is also hard to categorize whether Call of Duty as a shooter game or action-adventure game. If we categorize games and conduct the comparison, it will remain questionable if game category true-positively differentiates players' perceptions or actions. But we do recognize a future work possibility by first categorizing game genres systematically and then examining whether game genres influence players' perceptions and actions.

8 CONCLUSION

Player behavior, usually a combination of in-game language and avatar action, presents enormous challenges to behavior moderation systems. Penalties issued from moderation systems impact PX in profound ways but remain poorly understood. We thus conducted a survey study with 291 players to understand how they perceive and intend to behave around moderation systems in online multi-player games to obtain a clearer understanding of how to design more transparent and fairer punishment experiences. We found that compared to moderation notification, explanation plays a more critical role in improving players' perceived transparency and fairness of moderation. Also, compared to the perceived transparency, the perceived fairness more significantly affected players to adopt different coping strategies for punishments. As we found the importance of punishment explanation to perceived fairness, we emphasize the indirect role of explanation provision to support players in coping with punishments. These findings not only extend the understanding from prior moderation literature that frequently focuses on the social media context but also help frame moderation experience as part of player experience and rethink moderation design in online multiplayer game.

ACKNOWLEDGMENTS

This work is partially supported by NSF grant no. 2006854. We appreciate all anonymous reviewers' constructive feedback to make this work refined and improved. We also appreciate 291 players' participation in this survey study.

REFERENCES

[1] Leigh Achterbosch, Charlynn Miller, and Peter Vamplew. 2017. A taxonomy of griefer type by motivation in massively multiplayer online role-playing games. Behaviour & Information Technology 36, 8 (August 2017), 846–860. DOI:https://doi.org/10.1080/0144929X.2017.1306109

[2] Sonam Adinolf and Selen Turkay. 2018. Toxic Behaviors in Esports Games: Player Perceptions and Coping Strategies. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts - CHI PLAY '18 Extended Abstracts, ACM Press, New York, New York, USA, 365–372. DOI:https://doi.org/10.1145/3270316.3271545

[3] Aaron Alford. 2021. FIFA streamer loses his mind after being perma banned by EA live on Twitch. invenglobal.

[4] Anna Veronica Banchik. 2020. Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. New Media Soc (March 2020), 146144482091272. DOI:https://doi.org/10.1177/1461444820912724

[5] Nicole A. Beres, Julian Frommel, Elizabeth Reid, Regan L. Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. Conference on Human Factors in Computing Systems - Proceedings (May 2021). DOI:https://doi.org/10.1145/3411764.3445157

[6] Lauren Bergin. 2021. Epic permanently bans renowned Fortnite leaker HYPEX for incident 3 years ago. dexerto.com.

[7] Apurba Biswas. 2022. Free Fire Max: Garena permanently bans over 1.7M accounts for cheating in-game. insidesport.in.

[8] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from HeartMob. Proc ACM Hum Comput Interact 1, CSCW (November 2017), 1–19. DOI:https://doi.org/10.1145/3134659

[9] Kenneth A. Bollen and J. Scott Long. 1993. Testing Structural Equation Models. Sage.

[10] Jonathan E. Bone and Nichola J. Raihani. 2015. Human punishment is motivated by both a desire for revenge and a desire for equality. Evolution and Human Behavior 36, 4 (July 2015), 323–330. DOI:https://doi.org/10.1016/J.EVOLHUMBEHAV.2015.02.002

[11] Ragnhild Brøvig-Hanssen and Ellis Jones. 2021. Remix's retreat? Content moderation, copyright law and mashup music: New Media Soc (June 2021). DOI:https://doi.org/10.1177/14614448211026059

[12] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In ACM International Conference on Interactive Media Experiences. Association for Computing Machinery, New York, NY, USA, 61–72.

[13] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, ACM Press, New York, New York, USA. DOI:https://doi.org/http://dx.doi.org/10.1145/2818048.2819963

[14] Lars Thøger Christensen and George Cheney. 2015. Peering into Transparency: Challenging Ideals, Proxies, and Organizational Practices. Communication Theory 25, 1 (February 2015), 70–90. DOI:https://doi.org/10.1111/COMT.12052

[15] Rachel Ivy Clarke, Jin Ha Lee, and Neils Clark. 2017. Why Video Game Genres Fail: A Classificatory Analysis. Games Cult 12, 5 (July 2017), 445–465. DOI:https://doi.org/10.1177/1555412015591900

[16] Jason A. Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. Journal of Applied Psychology 86, 3 (June 2001), 386–400. Retrieved from https://psycnet.apa.org/buy/2001-06715-002

[17] Christine L. Cook. 2019. Between a Troll and a Hard Place: The Demand Framework's Answer to One of Gaming's Biggest Problems. Media Commun 7, 4 (December 2019), 176–185. DOI:https://doi.org/10.17645/MAC.V7I4.2347
[18] Julian Dibbell. 1994. A Rape in Cyberspace or How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database into a Society. Annu Surv Am Law (1994).
[19] Alexandra Durcikova and Peter Gray. 2014. How Knowledge Validation Processes Affect Knowledge Contribution. Journal of Management Information Systems 25, 4 (April 2014), 81–108. DOI:https://doi.org/10.2753/MIS0742-1222250403
[20] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk theories of social feeds. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2371–2382. DOI:https://doi.org/10.1145/2858036.2858494
[21] Pouyan Esmaeilzadeh. 2019. The Impacts of the Perceived Transparency of Privacy Policies and Trust in Providers for Building Trust in Health Information Exchange: Empirical Study. JMIR Med Inform 7, 4 (November 2019). DOI:https://doi.org/10.2196/14050
[22] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. Behav Res Methods 54, 4 (September 2021), 1643–1662. DOI:https://doi.org/10.3758/S13428-021-01694-3/TABLES/13
[23] Leandre R. Fabrigar, Robert C. MacCallum, Duane T. Wegener, and Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 4, 3 (September 1999), 272–299. DOI:https://doi.org/10.1037/1082-989X.4.3.272
[24] Fair Play Alliance. 2020. Disruption and Harms in Online Gaming Framework.
[25] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2020), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376293
[26] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. Proc ACM Hum Comput Interact 4, CSCW1 (May 2020). DOI:https://doi.org/10.1145/3392845
[27] Susan Folkman, Richard S. Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J. Gruen. 1986. Dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. J Pers Soc Psychol 50, 5 (1986).
[28] Chek Yang Foo. 2008. Grief Play Management. VDM Verlag.
[29] Claes Fornell and David F. Larcker. 1981. Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. Journal of Marketing Research 18, 3 (November 1981). DOI:https://doi.org/10.1177/002224378101800313
[30] Daniel Friedman. 2018. How Riot may have made it impossible to keep 'the most toxic League of Legends player' banned. Polygon.
[31] Janet Rausa Fuller. 2021. Report: Toxicity in gaming and the need for better moderation. PickFu.
[32] Phill Gagné and Gregory R. Hancock. 2010. Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. Multivariate Behav Res 41, 1 (2010), 65–83. DOI:https://doi.org/10.1207/S15327906MBR4101_5
[33] Marton Gergely and V. Srinivasan Chino Rao. 2017. Social desirability bias in software piracy research: Evidence from pilot studies. Proceedings of the 2016 12th International Conference on Innovations in Information Technology, IIT 2016 (March 2017). DOI:https://doi.org/10.1109/INNOVATIONS.2016.7880052
[34] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. New Media Soc 20, 12 (December 2018), 4492–4511. DOI:https://doi.org/10.1177/1461444818776611
[35] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. Soc Media Soc 5, 1 (January 2019), 205630511983258. DOI:https://doi.org/10.1177/2056305119832588
[36] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. Social Media and Society 8, 3 (July 2022). DOI:https://doi.org/10.1177/20563051221117552/ASSET/IMAGES/LARGE/10.1177_20563051221117552-FIG2.JPEG
[37] João Gonçalves, Ina Weber, Gina M. Masullo, Marisa Torres da Silva, and Joep Hofhuis. 2021. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. New Media Soc (July 2021). DOI:https://doi.org/10.1177/14614448211032310
[38] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data Soc 7, 1 (January 2020), 205395171989794. DOI:https://doi.org/10.1177/2053951719897945
[39] Austen Goslin. 2018. Tyler1 has officially been unbanned by Riot. The Rift Herald. Retrieved from https://www.riftherald.com/culture/2018/1/4/16850598/tyler1-unbanned-lol-reformed
[40] Douglas Gunzler, Tian Chen, Pan Wu, and Hui Zhang. 2013. Introduction to mediation analysis with structural equation modeling. Shanghai Arch Psychiatry 25, 6 (2013), 390. DOI:https://doi.org/10.3969/J.ISSN.1002-0829.2013.06.009
[41] David J. Harding, Jeffrey D. Morenoff, Anh P. Nguyen, Shawn D. Bushway, and Ingrid A. Binswanger. 2019. A natural experiment study of the effects of imprisonment on violence in the community. Nature Human Behaviour 2019 3:7 3, 7 (May 2019), 671–677. DOI:https://doi.org/10.1038/s41562-019-0604-8
[42] Stephanie Heintz and Effie Lai Chong Law. 2015. The game genre map: A revised game classification. CHI PLAY 2015 - Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (October 2015), 175–184. DOI:https://doi.org/10.1145/2793107.2793123
[43] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano de Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3476057
[44] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 1–33. DOI:https://doi.org/10.1145/3359294
[45] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Transactions on Computer-Human Interaction 26, 5 (July 2019), 1–35. DOI:https://doi.org/10.1145/3338243
[46] Shagun Jhaver, Christian Boylston, DIyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021), 30. DOI:https://doi.org/10.1145/3479525

[47] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. Proc ACM Hum Comput Interact 3, CSCW (2019). DOI:https://doi.org/10.1145/3359252
[48] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction 25, 2 (March 2018), 1–33. DOI:https://doi.org/10.1145/3185593
[49] Hua Jiang and Hongmei Shen. 2020. Toward a Relational Theory of Employee Engagement: Understanding Authenticity, Transparency, and Employee Behaviors: International Journal of Business Communication (September 2020). DOI:https://doi.org/10.1177/2329488420954236
[50] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-Based Online Communities on Discord. Proc. ACM Hum.-Comput. Interact. 3, CSCW (November 2019). DOI:https://doi.org/10.1145/3359157
[51] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. Proc ACM Hum Comput Interact 4, GROUP (January 2020), 1–35. DOI:https://doi.org/10.1145/3375197
[52] Jesper Juul. 2009. Fear of failing? the many meanings of difficulty in video games. The video game theory reader 2, (2009), 237–252.
[53] Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. Journal of Online Trust and Safety 1, 3 (August 2022). DOI:https://doi.org/10.54501/JOTS.V1I3.38
[54] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20), Association for Computing Machinery, New York, NY, USA, 81–92. DOI:https://doi.org/10.1145/3410404.3414243
[55] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. Proceedings of the Annual Symposium on Computer-Human Interaction in Play August (2020). DOI:https://doi.org/10.1145/3410404.3414243
[56] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3476075
[57] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: the Case of AI-Led Moderation. Proc ACM Hum Comput Interact 4, CSCW2 (October 2020), 1–27. DOI:https://doi.org/10.1145/3415173
[58] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA. Retrieved from https://doi.org/10.1145/3411764.3445279
[59] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. Proc ACM Hum Comput Interact 1, CSCW (2017), 62. DOI:https://doi.org/10.1145/3134697
[60] Haewoon Kwak and Jeremy Blackburn. 2015. Linguistic Analysis of Toxic Behavior in an Online Video Game. In Social Informatics: SocInfo 2014 International Workshops, Springer Verlag, 209–217. DOI:https://doi.org/10.1007/978-3-319-15168-7_26/COVER
[61] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, ACM Press, New York, New York, USA, 3739–3748. DOI:https://doi.org/10.1145/2702123.2702529
[62] Eric G. Lambert and Nancy L. Hogan. 2013. The Association of Distributive and Procedural Justice With Organizational Citizenship Behavior: Prison J 93, 3 (June 2013), 313–334. DOI:https://doi.org/10.1177/0032885513490491
[63] Min Kyung Lee, Anuraag Jain, Hae J.I.N. Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 26. DOI:https://doi.org/10.1145/3359284
[64] Jeffrey Lin. 2013. The Science Behind Shaping Player Behavior in Online Games. In Game Developers Conference.
[65] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–25. DOI:https://doi.org/10.1145/3449180
[66] Renkai Ma, Xinning Gui, and Yubo Kou. 2022. Esports Governance: An Analysis of Rule Enforcement in League of Legends. Proc ACM Hum Comput Interact (2022). DOI:https://doi.org/10.1145/3555541
[67] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. PACM on Human Computer Interaction 5, CSCW2 (2021), 1–26. DOI:https://doi.org/https://doi.org/10.1145/3479573
[68] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. Proc ACM Hum Comput Interact 6, CSCW2 (2022), 28. DOI:https://doi.org/10.1145/3555150
[69] Renkai Ma and Yubo Kou. 2022. "I am not a YouTuber who can make whatever video I want. I have to keep appeasing algorithms": Bureaucracy of Creator Moderation on YouTube. In Companion Computer Supported Co-operative Work and Social Computing (CSCW'22 Companion). Retrieved from https://doi.org/10.1145/3500868.3559445
[70] Connor Makar. 2022. Riot Games will soon start monitoring Valorant voice chat in an attempt to curb toxicity. vg247.
[71] Cass Marshall. 2022. Blizzard's new policy bans World of Warcraft boosting organizations. polygon.
[72] C. M. McBride, K. M. Emmons, and I. M. Lipkus. 2003. Understanding the potential of teachable moments: the case of smoking cessation. Health Educ Res 18, 2 (April 2003), 156–170. DOI:https://doi.org/10.1093/her/18.2.156
[73] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media Soc 20, 11 (2018), 4366–4383. DOI:https://doi.org/10.1177/1461444818773059
[74] Anton J. Nederhof. 1985. Methods of coping with social desirability bias: A review. Eur J Soc Psychol 15, 3 (July 1985), 263–280. DOI:https://doi.org/10.1002/EJSP.2420150303
[75] Brian P. Niehoff and Robert H. Moorman. 1993. Justice as a Mediator of the Relationship Between Methods of Monitoring and Organizational Citizenship Behavior. Academy of Management Journal 36, 3 (November 1993), 527–556. DOI:https://doi.org/10.5465/256591
[76] Stephany Nunneley. 2018. Fair Play Alliance formed by Blizzard, Epic, Twitch, Xbox to curb noxious behavior in online games. VG247.
[77] Fatimah Omar, Amiraa Ali Mansor, and Fatimah WATI Halim. 2011. The relationships between organizational justice, organizational citizenship behavior and job satisfaction. Pertanika Journal of Social Science and Humanities 19, S (2011), 115–121.

[78] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. J Behav Exp Finance 17, (March 2018), 22–27. DOI:https://doi.org/10.1016/J.JBEF.2017.12.004
[79] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. Proc ACM Hum Comput Interact 6, CSCW1 (April 2022). DOI:https://doi.org/10.1145/3512929
[80] Delroy L. Paulhus. 1991. Measurement and Control of Response Bias. Academic Press (1991), 17–59. DOI:https://doi.org/10.1016/B978-0-12-590241-0.50006-X
[81] Simon T Perrault and Weiyu Zhang. 2019. Effects of Moderation and Opinion Heterogeneity on Attitude towards the Online Deliberation Experience. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–12.
[82] Rajesh. 2016. Blizzard issues second massive ban to cheaters in Overwatch. gametransfers.com.
[83] Robert F. Scherer, David C. Luther, Frank A. Wiebe, and Janet S. Adams. 1988. Dimensionality of coping: Factor stability using the ways of coping questionnaire. Psychol Rep 62, 3 (August 1988), 763–770. DOI:https://doi.org/10.2466/pr0.1988.62.3.763
[84] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021), 33. DOI:https://doi.org/10.1145/3479512
[85] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–18. DOI:https://doi.org/10.1145/3449076
[86] Lawrence Scotti. 2021. League of Legends streamer permabanned for death threats in game chat. dexerto.com.
[87] 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 (CSCW '17), Association for Computing Machinery, New York, NY, USA, 111–125. DOI:https://doi.org/10.1145/2998181.2998277
[88] Sercan Şengün, Jung Soongyo, Bernard J. Jansen, Joni Salminen, and Peter Mawhorter. 2019. Analyzing hate speech toward players from the MENA in League of Legends. CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019) (2019), 1–6. DOI:https://doi.org/10.1145/3290607.3312924
[89] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. Comput Human Behav 108, (July 2020), 106343. DOI:https://doi.org/10.1016/j.chb.2020.106343
[90] Nick Statt. 2021. New Unity study shows just how toxic online gaming can be. protocol.
[91] 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. Int J Commun 13, (2019). Retrieved from https://ijoc.org/index.php/ijoc/article/view/9736
[92] Jeanna Sybert. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban: New Media Soc (February 2021). DOI:https://doi.org/10.1177/1461444821996715
[93] Lars Thøger Christensen. 2002. Corporate communication: The challenge of transparency. Corporate Communications: An International Journal 7, 3 (September 2002), 162–168. DOI:https://doi.org/10.1108/13563280210436772/FULL/XML
[94] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. (January 2022). DOI:https://doi.org/10.48550/arxiv.2201.06455
[95] April Tyack and Elisa D. Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. Association for Computing Machinery (ACM), 1–22. DOI:https://doi.org/10.1145/3313831.3376723
[96] Tom Tyler, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. Social media governance: can social media companies motivate voluntary rule following behavior among their users? J Exp Criminol 17, 1 (March 2021), 109–127. DOI:https://doi.org/10.1007/S11292-019-09392-Z/FIGURES/3
[97] George Ursachi, Ioana Alexandra Horodnic, and Adriana Zait. 2015. How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators. Procedia Economics and Finance 20, (January 2015), 679–686. DOI:https://doi.org/10.1016/S2212-5671(15)00123-9
[98] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. In Proceedings of the ACM on Human-Computer Interaction, Association for Computing Machinery, 1–22. DOI:https://doi.org/10.1145/3415238
[99] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021), 28. DOI:https://doi.org/10.1145/3476059
[100] Michael Wenzel, Tyler G. Okimoto, and Kate Cameron. 2012. Do Retributive and Restorative Justice Processes Address Different Symbolic Concerns? Crit Criminol 20, 1 (March 2012), 25–44. DOI:https://doi.org/10.1007/S10612-011-9147-7/TABLES/5
[101] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. 2008. Retributive and restorative justice. Law and Human Behavior 32, 375–389. DOI:https://doi.org/10.1007/s10979-007-9116-6
[102] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. 2010. Justice through consensus: Shared identity and the preference for a restorative notion of justice. Eur J Soc Psychol 40, 6 (October 2010), 909–930. DOI:https://doi.org/10.1002/EJSP.657
[103] Michael Wenzel and Ines Thielmann. 2006. Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. Soc Justice Res 19, 4 (December 2006), 450–470. DOI:https://doi.org/10.1007/S11211-006-0028-2/FIGURES/2
[104] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems 15, (April 2022). DOI:https://doi.org/10.1145/3491102.3517614
[105] Yajiong Xue, Huigang Liang, and Liansheng Wu. 2010. Punishment, Justice, and Compliance in Mandatory IT Settings. nformation Systems Research 22, 2 (February 2010), 400–414. DOI:https://doi.org/10.1287/ISRE.1090.0266
[106] Mustafa Mikdat Yildirim, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker. 2021. Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter. Perspectives on Politics (2021), 1–13. DOI:https://doi.org/10.1017/S1537592721002589
[107] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20), Association for Computing Machinery, New York, NY, USA, 365–378. DOI:https://doi.org/10.1145/3379337.3415858

[108] Procedural Justice. Yale Law School. Retrieved from https://law.yale.edu/justice-collaboratory/procedural-justice