Community Begins Where Moderation Ends

Peer Support and Its Implications for Community-Based Rehabilitation

Yubo Kou
College of Information Sciences and Technology, Pennsylvania State University, USA
yubokou@psu.edu

Renkai Ma
College of Information Sciences and Technology, Pennsylvania State University, USA
renkai@psu.edu

Zinan Zhang
College of Information Sciences and Technology, Pennsylvania State University, USA
zzinan@psu.edu

Yingfan Zhou
College of Information Sciences and Technology, Pennsylvania State University, USA,
yxz5975@psu.edu

Xinning Gui
College of Information Sciences and Technology, Pennsylvania State University, USA,
xinninggui@psu.edu

ABSTRACT

Moderation systems of online games often follow a retributive model inspired by real-world criminal justice, expecting that punishments can help users to reform behavior. However, decades of criminological research show that punishments alone do not work and call for a rehabilitative approach, such as community-based rehabilitation (CBR), to help offenders transform their minds and behavioral patterns. Motivated by this call, we explore how moderated users view punishments in a community context and how other community members respond in League of Legends (LoL), one of the largest online games. Specifically, we focus on how peer support is sought and provided on the /r/LeagueOfLegends subreddit, the largest LoL-related online community. Our content analysis of player discussions characterized the communication between moderated users and peers as informative, constructive, and reflexive. We highlight the importance of involving community in moderation systems and discuss implications for designing CBR mechanisms that could enhance moderation systems.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing**; • **Empirical studies in collaborative and social computing**;

KEYWORDS

community-based rehabilitation, moderation experience, online moderation

ACM Reference Format:
Yubo Kou, Renkai Ma, Zinan Zhang, Yingfan Zhou, and Xinning Gui. 2024. Community Begins Where Moderation Ends: Peer Support and Its Implications for Community-Based Rehabilitation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642675

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642675

1 INTRODUCTION

Online games, like other types of online platforms, rely on moderation systems to regulate their players, set community norms, and fend off toxic behaviors that violate game policies or community norms (e.g., [36, 54, 59, 93]). While moderation techniques and approaches are many, they commonly follow a retributive model inspired by real-world criminal justice [37, 86], where punishments such as content removal, restricted access, or account suspension [70] are issued to players, expecting them to reform behavior in the future. As a result, punished players oftentimes end up being left on their own to cope with their punishments.

While retributive justice has potential to help shape behavior in favorable directions (see [19, 42]), in the real world criminal justice, criminologists have already reflected on its limitations [24, 76, 78], such as how it ignores the social roots that cause crimes and punishments alone do not effectively reform offenders but put offenders in a further disadvantaged position, fueling their reoffending behavior. Similarly, moderation researchers have also observed how retributive moderation can fall short in several ways: Moderation decisions could be too vague to communicate either explanatory information or actionable suggestions to moderated users, who in turn could struggle to understand what platform rules they have violated and what to pay attention to in the future [43, 51, 68]. Moderation decisions could be exclusionary against newcomers [10]. When moderation decisions make a mistake, which is not uncommon [36], moderation systems oftentimes fails to provide a proper remedy [2, 32]. More specifically in the context of online game moderation, a recent survey study shows that multiplayer online game players do not necessarily understand why they are punished and desire more information and instructions on how they can improve behavior [70]. Clearly, retributive justice has its limits in online game moderation.

Given the pitfalls of the retributive justice model that overly relies on punishment, many criminologists have endorsed the *rehabilitative* approach, which emphasizes active interventions that could

work with offenders together to facilitate their behavioral transformations and social reintegration into community [24, 78, 83]. Generally speaking, while restorative justice attends to all stakeholders (e.g., victims, offenders, and bystanders), rehabilitative justice has a specific focus on helping offenders [24]. Particularly, criminologists have recognized the importance of community in the rehabilitative process [24, 90], or community-based rehabilitation (CBR), because ultimately it is the community that welcomes back the offenders and supports their successful social reintegration. Thus, it is important to explore how kinds of support a community could provide to its moderated users.

To address this question, we choose to focus on how players seek support after they are punished by video game moderation. Specifically, we unpack the question by exploring: 1) how punished players seek support from other community members, 2) how the player community responds, and 3) how such support seeking and giving correspond to each other. Importantly, these questions are not meant to identify tangible benefits a CBR approach provides to moderated users. However, by exploring what support resources already exist in the community context, our study could inform the design of future CBR approaches that could enhance moderation systems.

To answer these questions, we empirically investigated support seeking of punished players in the online community of League of Legends (LoL), one of the largest multiplayer online games today with 180 million monthly players [62]. We collected and analyzed data from the '/r/leagueoflegends' subreddit, where LoL users share their received moderation decisions and seek support from fellow LoL players. Through a content analysis [56] and then a co-occurrence analysis, we found eight ways moderated players engaged in the communication with fellow players about their moderation experiences ranging from their emotions to perceptions to information seeking about moderation decisions and designs. As fellow players responded to these moderated players, we found eight themes of it ranging from information provision about moderation to behavior suggestions to the endorsement of moderation decisions. Furthermore, we uncovered that three benefits of such peer support were conspicuous, including its informativeness, instructiveness, and criticism. Based on these empirical findings, we discuss the design implications for peer support mechanisms that could enhance existing moderation design.

The study makes two primary contributions to HCI and games research: First, with a descriptive nature, the study provides a systematic account of how punished users disclose their punishments, reflecting their experiences with and perceptions of punishments, as well as how other community members respond to those punished users, indicating commonly held values and norms within a community context. We see the focus on the particular competitive gaming context of LoL as a strength, as it enables us to develop nuanced, in-depth interpretations of LoL users' experiences. A systematic, descriptive account like this can provide an empirical basis for future research on cross-validation and cross-comparison, geared towards forming a broader interest in investigating the behavior and experiences of punished users in order to enhance moderation.

Second, by connecting empirical findings with criminological insights, the study articulates conceptual connections between existing community support resources and community-based rehabilitation. Not intended to demonstrate or measure the benefits of CBR, the study nonetheless offers ample design implications for leveraging CBR to enhance moderation systems.

2 RELATED WORK

In this section, we begin by synthesizing the moderation literature and discussing the importance of post-moderation support. We then present insights from criminology supporting the approach of community-based rehabilitation. We then engage with empirical and design studies of peer support to discuss how peer support can be a productive source of behavioral improvement and beneficial in enhancing moderation systems.

2.1 Moderation and the Need to Support Punished Players

Games research in human-computer interaction (HCI) has started to pay attention to toxic behaviors in recent years (e.g., [1, 5, 12, 50]). Here, toxic behavior refers broadly to behaviors that deviate from shared values and social norms in an online community, as well as rules and policies prescribed by the online platform, thus disrupting others' experience in the community. In the meantime, game researchers have also started to explore various moderation approaches that can curb in-game toxicity (e.g., [1, 54]). Moderation denotes a set of governance mechanisms (e.g., platform rules and technical means such as automated detection and removal of foul language) that seeks to foster cooperation and discourage disruptive behaviors in an online community [38]. While platforms, ranging from social media platforms to online games, have been rapidly expanding in the past years, modern moderation systems are oftentimes an afterthought in platform development [36]. As a result, moderation is often framed as a solution to the social problem of toxic behaviors. Various moderation approaches exist in terms of procedures and the tradeoff between automated techniques and human labor [84, 91], and oftentimes in combination on any particular platform [36]. A common moderation cycle that is familiar to many users is called ex-post reactive moderation, where a user is first flagged by automated tools or other users, and then the moderation system adjudicates on the flag and punishes the user, if convicted. In other words, the design of moderation has a clear-cut end — the issuance of a moderation decision.

Within the moderation literature, much attention has been paid to how moderation is carried out within the moderation cycle and before moderation ends. Scholars have researched individual moderation steps such as flagging [54], blocking [45], and adjudicating [27]. In addition, a recent body of work has paid attention to human moderators' practices [89, 98] and emotional experiences [80, 99]. Related to the scholarly attention to human moderators' emotional experiences, in recent years, media coverage has highlighted human moderators' poor work conditions and psychological struggles [49].

However, less is known about the social implications of moderation decisions for people impacted by a moderation decision. This remains a nascent research space within moderation literature, but existing research has collectively highlighted some gaps between

existing moderation systems' approaches and moderated users' reactions. First, punished users could be confused and struggle to understand why their behaviors have violated certain standards [55]. They do not necessarily have a firm grasp of community norms and official rules. They could perceive unfairness when comparing with other users' similar behaviors and moderation treatments [13]. While an appeal process has the potential to restore users' fairness perception, Vaccaro et al. conducted an experimental study [95] to investigate and reflect upon the appeal process of automated moderation decisions, revealing systematic failures of appeals, including power imbalance, isolated decision, and perceived illegitimacy. Second, users may benefit from explanations for moderation decisions. For example, Jhaver et al. tested the effect of providing explanations about content removal on Reddit and concluded that it might be worthwhile for moderation systems to provide explanations [43]. Kou and Gui found that users collectively generated explanations for why they were affected and how they might improve to avoid future punishments [52]. Third, punished users may seek to figure out moderation criteria and devise circumvention skills. For example, pro-eating disorder (pro-ED) users on Instagram would identify themselves as pro-ED without using related hashtags, which they knew would attract the attention of platform moderation [32, 34].

Taken collectively, the existing literature points to the limits of existing moderation approaches, as well as the need for design considerations to support punished users in the post-moderation phase. When punished users are left on their own to figure out how to deal with moderation penalties, penalties alone may not transform them into well-behaved or well-informed community members. Thus, this study seeks to address this gap by putting the human experience of moderation at the center.

2.2 Community-Based Rehabilitation: Insights from Criminology

In criminology, rehabilitation is defined as the process to "reduce recidivism among adults who have been convicted of an offense by improving their behaviors, skills, mental health, social functioning, and access to education and employment" [75]. It could happen in both correctional settings where a prisoner completes their sentence, or in community settings where individual offenders experience positive reinforcement through community-based activities [64] and volunteer activities [82]. The field of criminology has existed for hundreds of years, and yet it has taken researchers and policymakers decades to recognize the philosophy and benefits of rehabilitation over punishments [23].

Criminologists have identified a wide range of justifications for favoring the rehabilitative approach. There is wide recognition that offenders often come from disadvantaged backgrounds such as low socio-economic status, early childhood abuse, and lack of stable family conditions [24]. Merely punishing them does little to eliminate the strong impacts of these disadvantaged backgrounds on their behaviors. In fact, there is also sufficient empirical evidence suggesting that punishment is not as effective as rehabilitation in preventing offenders from recidivism [14, 24, 73]. In addition, rehabilitation embraces the premise that human beings, even offenders, can be viewed as "goal-seeking beings who act to accomplish meaningful aims" [25], and rehabilitation can encourage offenders to pursue human goods [25]. Besides, there is consistent public support for rehabilitation [46, 78]. Lastly, while punishment is ethically justified, the condition of imprisonment still puts significant, disproportionate burdens on offenders beyond losing freedom [48].

While the retributive approach often expects offenders to work on their own to reform behavior but ignores the broader contexts and obstacles that prevent offenders from doing so, rehabilitation promises several unique advantages in fostering offenders' behavioral transformations. The rehabilitative approach takes a constructivist view to highlight that offenses are learned behaviors and could be unlearned through offender reeducation [33]. Instead of viewing offenders as beyond redemption, the rehabilitative approach emphasizes identifying positive factors that can keep people away from crime or cause people to cease perpetrating crime [83], and transforming the experience of punishment into a positive factor that can promote growth and development [83]. Rich empirical evidence shows that successful rehabilitation programs have been able to promote self-determination and personal growth [25], bring social acceptance [83], and strengthen positive worldviews and self-identity [83]. As Bonta and Andrews argued clearly in favor of the rehabilitative approach, "Offenders are human beings, and the most powerful influence strategies available are cognitive-behavioral and cognitive social learning strategies. It matters little whether the problem is antisocial behavior, depression, smoking, overeating, or poor study habits—cognitive-behavioral treatments are often more effective than other forms of intervention" [6].

Taken together, criminologists have accumulated much knowledge about the rehabilitative approach. On one hand, such knowledge contains valuable insights for how moderation systems can incorporate this approach to more effectively help offenders transform their behaviors. On the other hand, we should be mindful of the distinctions between real world criminal justice and online moderation that prevent the direct application of knowledge from the former into the latter. This further necessitates further research to understand how and to what extent criminological knowledge about rehabilitation can be used to enhance offender treatment in today's moderation systems. Naturally, a human-centered design approach can (1) use empirical methods to understand what support resources already exist in the wild akin to CBR, (2) evaluate how effective these support resources are, and then (3) move to design interventions that incorporate CBR mechanisms. Following this logic, this study constitutes the first step and seeks to describe what support resources the community context already provides for punished offenders, as well as how offenders perceive and experience punishments, and how the community interacts with them. In so doing, we provide ample empirical evidence to inform future design efforts that implement CBR mechanisms in moderation systems.

2.3 Peer Support as Community-based Rehabilitation

Socialization refers to the process through which a community member acquires knowledge, skills, behaviors, and attitudes to participate in the community [16]. Online community researchers in HCI have highlighted the importance of socialization process in online community development and sustainment [21, 31]. In this study, punished players' socialization process refers to how

punished players could acquire knowledge about the community's behavioral standards and initiate behavioral improvement.

Peer support, where peers provide support to others, includes four primary categories [39]: (1) Emotional support through which empathy, love, trust, and caring is provided, (2) Instrumental support that manifests in tangible aid and services, (3) informational support, through which information, advice, and suggestions are provided to help problem solving, and (4) appraisal support that supports self-evaluation, such as constructive feedback and affirmation. Peer support has been valued as an important learning mechanism in various contexts. For instance, educational researchers found that peer support could help reinforce individual students' accountability and improve their teamwork performance in team-based learning (e.g., [7, 9, 17]). Organizational researchers point to peer support's role in facilitating mutual understanding and agreement and enhancing task performance (e.g., [26]). Importantly, online platforms have become an important venue for peer support, where fellow community or group members provide informational support to each other.

RQ1. How do punished players engage in communication about moderation with peers?

RQ2. What support do peers generate for those punished players in this communication?

RQ3. How do the ways that punished players engage with peers and peer support correspond with each other in this communication?

3 LEAGUE OF LEGENDS AND ITS MODERATION APPARATUS

League of Legends (LoL), developed and maintained by Riot Games (Riot), is a major eSports title and one of the largest multiplayer online games today, with 180 million players across several titles set in the LoL universe [62]. LoL's primary game mode is a 5-vs-5 match happening on the "Summoner's Rift" map, where two teams of five compete to destroy the opposing team's base. LoL also maintains a corresponding ranking and matchmaking system. The former calculates a player's skill and adjusts the player's rank based on each match's result, while the latter uses player ranks to generate a fair match. Riot has made numerous efforts in promoting the eSports culture around LoL (e.g., [92]) and centering and celebrating player ranks within the player community (e.g., [71]). In such a highly competitive, eSports culture, players easily experience negative emotions such as anxiety and frustration, which oftentimes lead to interpersonal aggression [53, 61].

Toxic behaviors, such as griefing, harassment, and trolling, are endemic to the online game culture and common in the LoL community [50, 58]. To curb player toxicity, Riot has innovated and maintained various endeavors to moderate player behavior (e.g., [55, 63]). At present, LoL's main moderation system follows a retributive model in an automated fashion. Players can use the 'report' button in the post-game lobby to flag toxic behaviors. The automated moderation system collects flags and adjudicates the flagged behaviors, making the final decision to pardon or punish the player in question.

Punishments issued by the system vary in forms and severity, ranging from an in-client warning message, chat restriction, seven-day account suspension, 14-day account suspension, to permanent account suspension. The punishment notification usually comes with details of why the player is punished, such as the player's in-game chatlogs that contain toxic languages. See Figure 1 for an example. By the time of this study, Riot does not provide other forms of information or support at the post-moderation phase.

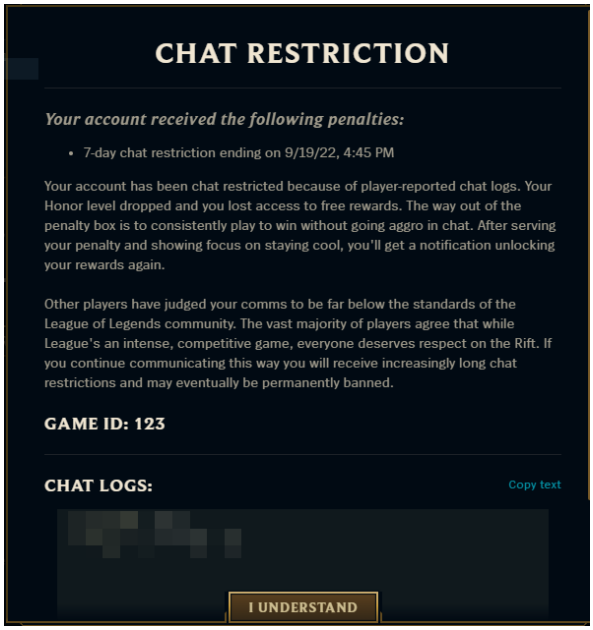

Figure 1: Chat Restriction in League of Legends (source: https://support-leagueoflegends.riotgames.com/hc/en-us/articles/205097293-In-Client-Penalty-Notifications-FAQ).

Besides this retributive method, Riot also utilizes positive reinforcement techniques. For instance, the Honor system allows players to command each other in one of several categories for the latter's positive behavior in game. Players could advance to higher honor levels and acquire corresponding in-game rewards. Another technique is positive messaging, where the game client displays a positive message when a game is loading [11].

4 METHODS

We will introduce how we collect and preprocess data from the 'r/leagueoflegends' subreddit for both content analysis (RQ1-2) and co-occurrence analysis (RQ3), as summarized in Figure 2.

4.1 Data Collection and Preparation

Through a four-step systematic effort, we identified threads discussing moderation experiences and peer support on the 'r/leagueoflegends' subreddit. Initially, in March 2022, three researchers compiled 26 keywords to search for threads related to players' moderation experiences. This compilation was done by (1) randomly searching and reading over 100 threads in 'r/leagueoflegends' to collect relevant terms or jargons, (2) examining related media reports (e.g., [29, 74]), and (3) reviewing prior literature on moderation experiences in both gaming (e.g., [52, 94]) and non-gaming contexts

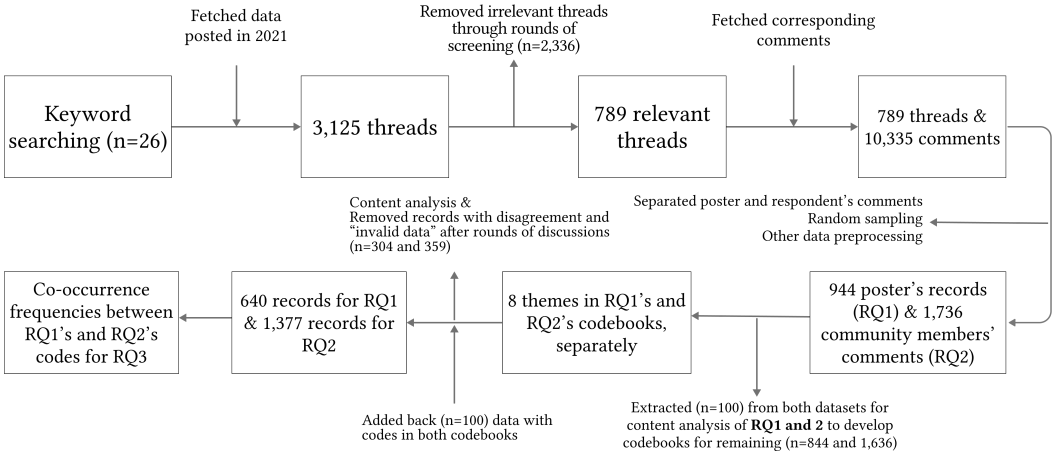

Figure 2: Method Flowchart. Please note that inward-pointing arrows represent adding records, while outward-pointing arrows represent extracting records or data preprocessing actions.

(e.g., [2, 97]). The 26 keywords included "got punishment", "received punishment", "banned", "warning", "7 day ban", "14 day ban", "got banned", "punish system", "AFK penalty", "was suspended", "account suspension", "permabanned", "perma ban", "permed", "permanent ban", "appeal ban", "temp ban", "week ban", "temporary ban", "chat restriction", "perma muting", "chat restricted", "game restriction", "rank restricted", "rewards are withheld" and "leaver penalty."

Next, in April 2022, we retrieved 3,125 threads from 'r/leagueoflegends' containing any of these keywords, using Pushshift Reddit API [3] and Pushshift Multithread API Wrapper (PMAW) on Python 3. We set the search time frame for threads between January 1, 2021, and December 31, 2021, covering the full season eleven (2021) of LoL.

Subsequently, three researchers screened these 3,125 threads to identify those most relevant to players' moderation experiences, specifically looking for explicit mentions of particular types of punishment (e.g., permanent ban and chat restriction). Given the large scale of this dataset, we randomly distributed the threads to ensure that each was screened by two researchers. The observed agreement rates and Cohen's kappa for each pair of researchers were (1) 89.33%, $\kappa = 0.74$, (2) 95.76%, $\kappa = 0.91$, and (3) 93.48%, $\kappa = 0.87$. All Cohen's kappa values indicated substantial (0.61-0.80) and nearly perfect (0.81-1.00) agreement [77]. Then, three researchers held rounds of meetings to resolve all the disagreements. This step led to a total of 1,204 threads that described players' moderation experiences. After eliminating threads with fewer than three comments to ensure meaningful engagement, as recommended by prior research (e.g., [28]), 789 threads remained, comprising 10,335 comments.

Finally, from the 789 threads, we further created two datasets: one comprising 2,916 entries (posts and comments) from punished players and another with 8,208 from community members. For the latter, we excluded 2,655 comments deemed short and non-informative (e.g., "XD" and "Yea"), in line with data preprocessing strategies used in prior work (e.g., [67]). Our datasets thus included 2,916 entries from punished players and 5,553 comments from community members. Due to the sizeable qualitative coding, we randomly sampled [47] approximately 31% from each, resulting in 944 records from punished players and 1,736 community comments for further analysis.

4.2 First Stage: Content Analysis and Codebook Development for RQ1-2

We conducted a content analysis [56] on both datasets. For each dataset, three researchers coded 100 random records from the two datasets (i.e., two sheets x three coders) in an inductive approach for RQ1 and RQ2. They assigned codes to concepts expressed in the punished players' records and community comments. Weekly discussions with two additional researchers helped to refine code definitions, consolidate similar codes, eliminate "invalid data" that contained little information or could answer RQs, address disparities between codes and data records, and identify higher-level themes emerging from codes. This collaborative effort reached 100% agreement on all the codes generated for the sample data, resulting in two codebooks. During the coding process, we distinguished between responses to the initial post and subsequent interactions to accurately capture dialogue dynamics.

Subsequently, three researchers individually applied the two initial codebooks to the remaining 844 punished players' records (RQ1's dataset) and 1,636 community comments (RQ2's dataset). They also held weekly meetings to discuss whether new codes emerged and needed to be added to the existing codebooks. This process finalized the two codebooks, with one for RQ1 including eight themes and 14 subthemes/codes (see Table 1) and the other for RQ2 including eight themes and 17 subthemes/codes (see Table 2). The Krippendorff's alpha values (i.e., inter-rater reliability) for codes on RQ1's (n=844) and RQ2's remaining dataset (n=1636) across three researchers were $\alpha = 0.73$ and $\alpha = 0.86$, indicating good agreement [57]. During this phase, we also filtered out data records that were irrelevant or uninformative to RQ1 and RQ2. For example, RQ2's dataset included comments such as "Mods (in r/leagueoflegends) should remove posts like that they are so bad," and "Great, nobody cares, this post is getting removed anyways."

These comments primarily addressing the subreddit itself rather than LoL, along with other uninformative ones like "the number of games won't make much difference if you play every day," were consequently excluded. The process resulted in 640 pertinent punished players' records and 1,377 community comments for RQ2's data.

4.3 Second Stage: Co-occurrence Analysis for RQ3

Based on the content analysis results from the first stage, we present a co-occurrence analysis for RQ3, examining how punished players' communication engagement corresponds with the emergence of peer support types. Unlike traditional, inferential statistical methods like Pearson's correlation or chi-square tests, this analysis can help understand patterns of relationships between codes in RQ1's dataset and RQ2's, within the same threads, rather than arguing for generalizability of peer support. Methodologically, we considered each raw data point such as a user comment or post in both dataset as an distinct instance, recognizing that a single data point could embody several themes simultaneously. This approach allowed us to precisely track and analyze the patterns in which different themes appeared together within one punished player's record or community member's comment.

To address potential confounding factors like the varying base frequencies of codes, we developed the 'Standardized Co-Occurrence Rate,' p(code A ∩code Bcode A)*p(code A ∩code B|code B), after we assigned a binary value ('1' for presence, '0' for absence) to each code in our datasets. This formula quantifies how likely it is for two specific codes (e.g., code A from RQ1's dataset and code B from RQ2's dataset) to appear together, considering their individual occurrence probabilities. Essentially, it assesses the combined occurrence of two codes while factoring in their separate probabilities. Due to a highly skewed distribution in our data (e.g., the theme occurrence of negative emotional expression is lower than others), we applied log transformation and standardization to these co-occurrence rates to rectify the skewness and enhance our visualizations (see Table 3). So, we added the "Standardized" before this Co-Occurrence Rate. We recognize that pointwise mutual information (PMI) is a commonly accepted standard for such analysis. However, PMI, in its typical form, might not adequately adjust for the skewed distribution of code frequencies (e.g., low or zero combined frequencies) in our dataset. Our 'Standardized Co-Occurrence Rate' was thus tailored to our data's unique characteristics.

This approach did lead to the presence of negative values, which reflect a lower-than-average co-occurrence rate between two codes relative to the dataset as a whole. These negative values do not indicate negative probabilities but rather signify less frequent co-occurrences than other code combinations. This nuanced interpretation is essential for understanding the dynamics of peer support in our study. We log-transformed and standardized the co-occurrence rates for a clearer visualization of these complex relationships, as we called the rates 'Standardized Co-Occurrence Rate.' We conducted the computation using Python and visualized the results through Excel.

5 FINDINGS

We identified eight typical themes in communications initiated and engaged by punished players with fellow players, and further eight primary types of peer support that fellow players provided to these punished players. Then, we will discuss peer support's benefits/utilities in the context of communication among players. Please note that we used "they/them/their" pronouns for interpreting qualitative data as we cannot assume players' gender. Additionally, due to the nature of content analysis, each record or quotation might have been assigned multiple codes/themes.

5.1 RQ1: Player Moderation Experience Sharing

We identified eight typical ways in which punished players engaged in conversations about moderation decisions with other players, as detailed from Sections 5.1.1 to 5.1.8 and summarized in Table 1.

5.1.1 Negative Emotional Expression. After experiencing moderation, players expressed various negative emotions such as anger and frustration. For example, one player posted a thread stating:

> I just got suspended for 14 days for the first time. (. . .) But tbh I got 11k ping when my net used to be better, like yesterday, and someone just starts like, "report because he's having bad ping" I got mad and furious because of that message and ended up losing the game and suspended. Even if the champs gank[1] on me when I'm almost taking down an enemy champ, then call me "useless," and they even didn't come.

"11k ping" refers to the player's high network latency which could hinder their in-game performance. The punished player in the above case was complained by other players for such high network latency. However, the punished player disagreed and believed they were helpful to the team by "almost taking down an enemy champ." Thus, they expressed strong negative emotions using words such as "mad and furious."

Players also expressed frustrations related to how they met uncooperative teammates and received moderation punishments afterwards. For example, a player wrote, "*It's just so incredibly frustrating getting flamed when I have 2 afks on my team.*" In this example, the punished player voiced their frustrations with teammates who went AFK [away from keyboard] in game.

Additionally, some players expressed sadness about moderation decisions, such as "*It's really sad to see my account banned forever.*"

5.1.2 Perceived Unfairness of Punishments. Echoing what much prior moderation work has uncovered, that users often perceived unfairness in the moderation decisions they received (e.g., [40, 95]), a significant theme in our study also highlights such perceived unfairness of punishments in games. For example, a player wrote, "*I got banned for 14 days unfairly cause all I said, 'I sure love losing before I get to play.'*" This player argued that their chat content was neither harmful nor severe enough to justify a 14-day game ban.

Other players further developed perceived unfairness from perceived inconsistency in moderation decisions across players. For example, a player wrote:

[1]"Gank" describes how players collaborate to ambush opponent players.

Table 1: Eight ways in which moderated players engage in communication with other players. Each percentage is the ratio of the number of occurrences of each subtheme/code to the total number of records (N=640).

	Theme(#, # of codes/640)	**Definition**	**Subtheme/code**(# of subthemes/640) Note that a few themes only have one subtheme.
1	Negative emotional expression (19, 2.97%)	Players expressed negative emotions, such as anger and frustration, when sharing their punishment.	Negative emotional expression (2.97%)
2	Perceived unfairness in punishment (177, 27.66%)	Players perceived the punishments they experienced as unfair.	Perceived unfairness of punishment (27.66%)
3	Details of punishment context (222, 34.69%)	Players described the context where they experienced punishments.	Details of punishment context (34.69%)
4	Introspection (98, 15.31%)	Players explained their rationales, mental processes, or reflections on their past or planned future behaviors, given their moderation experiences.	Explaining past behavior (7.66%) Proposing future behavior (5.00%) Reflecting on past behaviors (2.66%)
5	Complaints about moderation design (107, 16.72%)	Players expressed complaints against moderation or game.	Complaints about the lack of support (3.44%) Complaints about platform or platform designs (13.28%)
6	Attribution to toxic culture (38, 5.94%)	Players attributed their toxic behavior to the toxic culture.	Attribution to toxic culture (5.94%)
7	Disagreement with community members (31, 4.84%)	In conversations regarding moderation experience, the punished players disagreed with other players.	Disagreement with community members (4.84%)
8	Information-seeking about punishments (139, 21.72%)	Players posted threads to seek information about punishment in terms of how it works, reasons for it, how to avoid or resolve it, and its impacts.	General information-seeking about punishments (14.38%) Information-seeking on avoiding future punishments (2.19%) Information-seeking on resolving prior punishments (5.16%)

> I shouldn't get banned because I had one bad match where I got abused and told to kill myself all I did was defend myself, yet because of Riot's archaic system, I'm the one getting punished. I've lost nearly a decade's worth of progress and all my money, and that is truly unfair.

This moderated player developed the perceived unfairness based on two reasons. First, they believed that moderation system was not advanced enough to identify and punish the actual toxic players, leading to their wrongful punishment. Thus, they felt an inequality in moderation decisions among players. Second, because of this moderation decision, the punished player lost their efforts and money invested in their player account, exacerbating their sense of unfairness.

Resonating with prior work on how players misuse flagging/reporting system to intensify in-game toxicity [54], our findings suggest that moderated players found it unfair when other players leveraged such flagging system against them. A player wrote, "*This was one bad game where 4 people reported me. That's just not fair.*"

Even after initiating an appeal and interacting with the human support team (e.g., moderators), their perceived unfairness could still remain. One player wrote:

> So, I have got temp[orarily] banned a month ago for calling someone on my team an ape[2], but I submit video evidence of someone on my team flashing[3] directly into the enemy team and not playing with the team at all. I'm just curious how the support team can tell me, "Well, it's a different process, and it is harder to tell if someone is intentionally griefing or is just having a rough game."

In this case, the punished player perceived the temporary ban as unfair because "flashing directly into the enemy" was a clear indication of wasting valuable resources, which contradicted the support team's argument about the challenge in recognizing intentional toxicity.

5.1.3 Details of Punishment Context. When punished players bring up the moderation decisions they had received, they often provided contextual information to render their experience more vivid and to justify their behaviors. For example, a punished player shared:

> "(. . .) But allow me to give you context. Vi[4] said I solo lost the game after she went and died due to a very

[2]"Ape" is a jargon to accuse other players of poor performance.
[3]"Flash" is an ability that allows a player's character to move in a short distance within a fleeting time. It is considered a valuable, scarce resource since it is commonly used to escape from enemy attacks, and the cool-down time for Flash is longer than most other abilities.
[4]"Vi," short for Violet, is a character in game.

> bad gank. (. . .) She was sitting under my turret all game, and jungle refused to gank whilst blaming me for the loss. Everyone else in the game was calling me bronze and trash, saying I solo lost the game. But I get banned for what?

In this case, the punished player described that Vi's unsuccessful gank benefited the enemy team. Furthermore, Vi stayed in this lane which belonged to the punished player, meaning that Vi split the resources and thus hindered the punished player's growth. The punished player thus attributed the loss of the lane to Vi. However, the team blamed the punished player instead. Thus, the punished player provided such contextual information to justify their actions and question the legitimacy of the punishment.

Sometimes, players would further elaborate on in-game conversations in game to argue that their language use was less toxic than that of others. For example, a player wrote, "*In games where I got told to kill myself, but I just said they are losers.*"

Moreover, some players who received warnings for being AFK cited offline, external difficulties (e.g., poor internet connections) to justify their AFK behavior. A player explained, "*For some reason, my screen kept going black at times in game, too, so that made it hard, but I still managed to play.*"

5.1.4 Introspection. We observed many punished players explaining their rationales, mental processes, or reflections behind their past or intended future behaviors. For instance, a punished player shared, "*It's hard not to type back when you're getting flamed when playing 10 hrs. every day.*" This example showed that the player admitted the legitimacy of their punishment and explained that they reacted toxically due to encountering other toxic players.

Punished players also recounted instances where they were provoked by other toxic players. They described their mental struggles, highlighting the difficulty in adhering to platform rules or avoiding toxic behavior:

> I couldn't control myself under them (other players) spam laughing[5] and my team dancing in base[6], and that's my bad.

In this case, the enemy team acted provocatively through spam laughing. The punished players' teammates were also disengaged in the gameplay. The punished player rationalized their toxic behavior as a loss of control, attributing it to others' toxic actions. They emphasized that these actions pushed them to a breaking point, leading to their own toxic response.

Related to the above case, punished players oftentimes actively reflected on their past behaviors and shared their reflections with fellow players. For example, a player remarked, "*I have reformed so much from my actually toxic self. I have learned to laugh at overtly toxic people.*" This punished player acknowledged their previous toxicity. They began to mentally distance themselves from such behavior and avoid reacting toxically when encountering other toxic players.

Punished players even detailed future behavioral strategies they planned to implement to avoid punishments. For example, a player wrote, "*I have that (chat) off too, but if it means me getting reported for playing bad can get me permabanned, then I'll take a break from the league. I don't want to lose my main account.*" This player had disabled the chat function in game and was even considering pausing their gameplay to prevent more severe punishments.

5.1.5 Complaints about Moderation Design. Some moderated players complained about the game and its moderation design. For example, a player complained about the inconsistency in how LoL's moderation system makes decisions:

> I have called people degenerates in many games, but this is the first time I've ever gotten punished for it. I deserved to be punished, yes, but if I'm going to get punished for saying the word, I should get punished for it every single time I say it.... I want consistency cause without consistency. The entire system is broken.

The player in this case perceived inconsistencies in moderation design. The punished player believed that effective moderation design should consistently apply.

In addition to voicing complaints about the moderation design, players also wrote about the lack of support from the game platform. As a player wrote, "*Do you know if there any steps, even legal ones, which I can take to fight for my rights? I invested a lot of time/cash into the account just to hit the wall called customer support.*" This player was exploring ways to address moderation issues concerning their account.

5.1.6 Attribution to Toxic Culture. Punished players also attributed their toxic deeds to the game culture. For example, a punished player described LoL itself is generally toxic, as they wrote: "*I'm not the problem. I play cs go*[7] *and I am competitive there; I still have tons of fun playing it, but the game (LoL) is fucked, not me.*"

Resonating with previous research's findings that toxicity breeds toxicity [50], our findings also indicate similar cases. For example, a moderated player wrote: "*I try to be friendly as possible but only then encounter people that are bad but judging me cause I can't do anything about a fed*[8] *enemy. I get toxic because of toxic people basically.*" This punished player complained that their teammates were unfriendly and unhelpful, leading them to adopt similar behaviors.

Some punished players even complained about the toxicity they experienced in LoL with sarcasm. A player wrote: "*I love league of legends and it's so awesome how your teammates report you and talk shit every game while greifing and inting themselves. Wow what an experience my 2 games have been today.*" Such sarcasm showed that negative behaviors this player encountered also negatively affected them, which might further perpetuate and reinforce a toxic culture.

Sometimes, punished players intentionally adopt toxic behavior to retaliate against others' toxicity. For example, a player stated, "*I truly think the best thing is just go and feed too, pretending to be bad, so the game ends quickly and nobody can tell you anything.*" Thus, the punished player suggested that adopting toxic behavior was a tactic to quickly end a game and thus reduce interaction with toxic players.

[5]"Spam laughing" refers to players sending laughing emotes to all players in the game, sometimes repeatedly, which is commonly considered as scoffing.
[6]"Dancing in base" refers to one side of players letting their characters dance in their base instead of defending against the enemies.
[7]Counter-Strike: Global Offensive is a multiplayer tactical first-person shooter game.
[8]"Fed" means a player has already scored many kills and become powerful in game, and any player from the enemy team can hardly beat them.

Due to the prevalence of toxicity, some punished players prevent themselves from being more toxic through game features. For example, a punished player disabled the chat function to avoid producing more toxic speech in LoL: "*The best way (of not being toxic) is to mute the chat. It is not the solution to the problem but spares one's nerves immensely.*"

5.1.7 Disagreement with Community Members. A small theme in our findings illustrates how punished players disagreed with responses or feedback that fellow players offered. For example, after posting their chat log (i.e., evidence), a punished player replied to a fellow player: "*Read the logs. Didn't type anything to anyone that wasn't me repeating what they said to me. Nothing racist.*" This player defended themselves, clarifying that they did not behave inappropriately.

Due to disagreements with fellow players, some punished players even threatened to leave the LoL community. A player wrote, "*You didn't read. Nothing I said was bad or offensive. Ok, bye, crazy community.*" Such comments suggest that some punished players felt unsupported by their peers in the LOL community.

5.1.8 Information-seeking about Punishments. A considerable number of punished players sought information about their punishments, such as their operating mechanisms, impacts, or potential remedies. For example, a punished player asked:

> It says I would also lose rewards and free bonuses also. What are those? I have also seen some stuff about a reform period. How long is the reform period?

The player's questions indicate a need for clarification on the specifics of their punishment, suggesting that the moderation design of LOL may not be adequately informative or effective in helping players understand an learn from their sanctions.

In addition to seeking knowledge about punishments, players also sought information on resolving their perceived moderation issues. A player experiencing a permanent ban wrote:

> I have a perma ban, so it must have been a serious 3rd party app. I checked my play history; about a month ago, someone played 3 games where it was like 15-0 [killing 15 enemies and getting killed zero times] on Vel'Koz[9], who I never play, and I am still a new player I couldn't get that KDA[10]. I messaged support all this, and they said they didn't find any suspicious logging activity. How is there no suspicious behavior; surely the hacker logged in from another IP address as well. Is there really nothing else I can do?

A score of 15-0 on a champion that the player had never played was an improbably high KDA for a beginner like the punished player to achieve. The punished player thus suspected that their account had been compromised and the hacker used a 3rd party app[11] to gain an unfair advantage in the game, leading to the permanent ban. Despite their suspicions, the support team could not help resolve the perceived moderation issue. The player had to seek information and guidance from other community members.

[9]"Vel'Koz" is a powerful character in LoL.
[10]"KDA" refers to the kill-to-death ratio that an individual champion has on average.
[11]https://support-leagueoflegends.riotgames.com/hc/en-us/articles/225266848-Third-Party-Applications

Similar to how social media users make sense of moderation decision and collectively learn about how to prevent it from happening to them again (e.g., [32, 34]), LoL players do the same. For example, a player who received chat restriction asked: "*I really want to stop before getting a permaban ban. Can you guys share your best tips on how you manage to resist trolling every game you're in when it's so goddamn hilarious?*" This player was actively seeking advice from peers on maintaining composure in the games' challenging and often toxic environment.

5.2 RQ2: Peer Support about Moderation

Analyzing the threads and comments that punished players posted to engage in conversations with peers about their moderation experiences, we found eight primary themes of peer support offered by fellow players, as seen in Table 2.

5.2.1 Future Behavior Suggestion. A primary theme in peer support is the provision of detailed behavioral advice to punished players. Sometimes, players suggested punished players contact the LOL support team. A player wrote, "*If you actually are innocent, I would make a Riot support ticket and see what they might say.*"

Oftentimes, players directly provided specific suggestions for future in-game behaviors to help punished players avoid future issues. For instance, "*If you don't type a novel in chat, you won't get chat restricted*" implies that punished players could type less in the future. Similarly, "*The best advice is if the internet isn't holding out, don't play ranked*" suggests that players should ensure a stable internet connection before engaging in ranked games.

We also observed that players' advice extended beyond the gaming context. For example, when a punished player explained that they had to be away from keyboard to help a family member and later received a warning, a fellow player suggested ways of avoiding future punishments:

> Inform your family that when you queue up for a game, you are making a commitment to finishing the game you started, regardless of the length of time it takes.

This above example shows that fellow players' feedback was not limited to in-game behavioral suggestions but also considered offline contexts in order to decrease punished players' future possibility of violating platform policies.

Beyond the purpose of avoiding punishments, the suggestions from peers also aimed to help punished players avoid experiencing toxicity. For example, a player stated, "*I only play this game now with friends and never read chat; if someone spam (with) clicks, I mute them" to avoid others' toxicity or "find a premade (e.g., friends) to play with that reduces the chance of [being] toxic.*"

Additionally, many peers offered detailed ways for punished players to manage in-game emotional state. For example, a player responded with tips for detaching from a negative emotional state, "*Usually, I'll either simply take a break, play a different game that isn't competitive, or take a little walk. The fresh air is nice to cool down after really bad ones, I find.*"

Beyond emotional management, peer feedback often urged punished players to introspect about their past behavior. For example, a player directly called out the toxic action shown in their chatlog, "*Let's be real, you're saying 'dog' instead of a proper swear word or*

Table 2: Eight themes of peer support that fellow players offered to punished players. Each percentage is the ratio of the number of occurrences of each code to the total number of records (N=1,377).

	Theme(#, # of codes/1,377)	**Definition**	**Subtheme/code**(# of codes/1,377)
1	Future behavior suggestion (389, 28.25%)	Players suggest punished players adopt certain behaviors for a better gaming experience, benefiting punished players themselves and others.	Advice on dealing with the toxic environment (1.45%) Suggesting punished players to reflect on past behaviors (3.20%) Directing punished players to contact platform support (5.08%) Other general future behavior suggestions (18.52%)
2	More evidence enquiry (55, 3.99%)	Players asked punished players for more contextual information about their punishments.	More evidence enquiry (3.99%)
3	Resonance or empathy on moderation experience (204, 14.81%)	Players shared their own moderation experiences or related to other players' experiences.	(Similar) experience sharing (11.69%) Expressing support or empathy to punished players (3.12%)
4	Complaints about moderation design (98, 7.12%)	Players complained about a lack of support or about the platform or its designs.	Complaints about the lack of support (0.87%) Complaints about platform or platform designs (6.25%)
5	Information provision about punishments (346, 25.13%)	Players offered information, in a confirmatory way, about moderation (e.g., punishment designs or impacts, news) or guesswork about punishments (e.g., punishment reasons or designs)	Guesswork on punishment reasons (8.06%) Guesswork provision around punishments (2.83%) Other general information provision about punishments (12.85%) Internet sources about moderation sharing (1.38%)
6	Support of moderation decisions (253, 18.37%)	Players affirmed the necessity or legitimacy of punishment on punished players' moderation experiences.	Support of moderation decisions (18.37%)
7	Attribution to toxicity culture (100, 7.26%)	Players expressed that the player culture could be toxic.	Attribution to toxicity culture (7.26%)
8	Disagreement with community members (84, 6.10%)	In conversations regarding moderation experience, players expressed disagreement with each other.	Disagreement with community members (6.10%)

racial slur. You're actively insulting your team and asking for reports. But you don't think you're doing anything incorrectly?"

5.2.2 More Evidence Enquiry. In response to accounts of punishments shared by players, some peers responded by asking for additional contextual information to ascertain whether the punished players were innocent or guilty. For example, they often asked for chatlogs that lead to a moderation decision. For instance, one comment was: "*Permabans are not given lightly, so unless you show us the logs, most people are just going to assume you said something bad enough to get permabanned.*"

In the absence of such evidence, peers remained skeptical. For example, here is a typical response, "*Going to need to see that ban card. No one on Reddit is that gully to believe your half-truth story.*" Here the "ban card" is an idiomatic expression for the in-client penalty notification that details what triggers a player's punishment [20].

5.2.3 Resonance or Empathy on Moderation Experience. Many players responded to punished players' experiences with resonance or empathy by sharing similar moderation experiences or offering emotional support. Such support demonstrated how punished users could potentially recover from the negative impacts of moderation through peer support (e.g., [32]). For example, one player empathized with a punished player by sharing their own challenging experiences:

> As far as I can tell, they won't fix anything. I'm getting banned on every single account I'm login in to, and they won't tell me a reason why. I got like 7 accounts banned for playing a single game - including my friend's. After sending a ticket, they just politely ignored me.

Players also shared successful experiences in addressing moderation issues. For instance, a player shared:

> Ask them (the support team) to clearly explain, step by step, what you did wrong since you want to understand. That's what I did, and then they realized

> the ban was bogus and unbanned me after I explained that the gameplay was not inting[12].

Here, the fellow player shared their experiences to encourage the punished player to communicating with the support team and establish their innocence.

Specifically, emotional support was apparent when fellow players shared their similar moderation experiences. For instance, players wrote comments like, "*Bro, I've had this happen to me for the past 3 months. (. . .) I feel your pain" and "I feel you man... I had my 5 years old account permabanned; I created another one, but it doesn't feel the same.*" Through such expressions, players conveyed empathy.

5.2.4 Complaints about Moderation Design. Along with empathy with punished players, the responses from fellow players frequently included complaints. These complaints echoed those of punished players, reflecting perceptions of inadequate support from the game company and flaws in moderation design. For example, a player shared:

> Riot's banning system is a mess. I had a similar ban; you can probably try and get an appeal done. Riot slapped me with a 28 day ban for an issue I couldn't fix. Riot even said it wasn't my issue that it was a server issue. (They didn't reverse it in my instance)

This player recounted an experience with moderation similar to that of the punished player. They criticized the moderation system for its inability to make accurate judgements and rectify its errors. Players also expressed dissatisfaction with the prolonged time required for resolving appeals on moderation decisions, as one remarked, "*I put in a ticket but says it can take up to 30 days, and it's just so disheartening.*" A player summarized, "*They got no interest in working through appeals.*" Such comments echo HCI researchers' calls for improved contestability in moderation systems (e.g., [66]).

5.2.5 Information Provision about Punishments. A great number of fellow players responded to punished players with information about the moderation system's decision-making. Sometimes, the information could be straightforward, such as news and Riot's policies. Like a player attached, "*https://www.riotgames.com/en/terms-of-service, number 8: Spamming. You said more in that game than I'll say in a month. You broke the TOS (Term of Services).*"

Other times, fellow players did not provide sources like news or TOS, but supplied information in a reductionist way. For example, a player explained how players would receive a permanent ban: "*A player's behavior is serious (. . .) [and] consistent enough to demonstrate that they have no interest in reform, and therefore and extra chances are going to be meaningless.*" Even though the player in this example did not specify the severity or frequency of problematic behaviors that led to a permanent ban, they insisted that severity and frequency were sufficient factors to warrant a permanent ban.

From a different angle, another player explained how reoffending could compound an existing punishment. For example, a player responded, "*If you are a player with a recent 14-day ban and you offend again, the only metric of severity that matters is 'bad enough to punish yes/no.' If the answer is yes, the next tier is perma-ban.*"

[12]Inting, or 'intentional feeding,' involves dying on purpose, and thus contributing to the enemy team's growth.

In addition, fellow players also actively offered guesswork about moderation. For instance, a player responded to another player who suspected a punishment:

> It won't actually be a punishment if you see the message, it's just the automated system noticing you weren't moving. Would love it if someone confirmed this or have a source; I am not 100% sure.

This example illustrates a player attempting to decipher the workings of algorithmic moderation. They suggested that it might be a mere system alert rather than a formal punishment.

Interestingly, as algorithmic moderation requires the coordination and collaboration of human workers [41], players also speculated on how Riot's support team handles moderation cases, factoring in the company's operational hours. A player wrote, "*Riot headquarters is in California. (. . .) I don't know exactly when you will get unbanned, but I would say minimum 8 more hours.*"

Such guesswork provision also involved sense-making on why players were punished. For instance, a player wrote:

> You can absolutely get account banned by reports in a free game by ruining others' ranked experiences by playing in a way that doesn't optimize winning in a game mode (ranked[13]) that people play to optimize winning.

The player inferred that the punished player did not play well, leading to other players reporting them.

5.2.6 Support of Moderation Decisions. A number of players endorsed the necessity or legitimacy of punishment on punished players. For example, a fellow player directly pointed out a punished player's toxic behaviors and supported the punishment, writing, "*telling people to report him is not attacking? Calling him a troll is not attacking? Asking if he is 12 is not attacking? You earned the chat restriction.*"

Some fellow players even hoped for severer punishments to be issued to the punished players. One wrote, "*I feel like leavers should deserve a little more of a beating.*"

In addition to straightforward affirmations that penalized players deserved their punishments, some players employed sarcasm to convey this viewpoint. For instance, in response to a player complaining about the unfairness of their punishment, another retorted, "*Calm down dude, you're not the joker. You're just a joke.*"

5.2.7 Attribution to Toxicity Culture. Similar to what punished players described about the toxic game culture, fellow players' feedback often frequently emphasized the pervasiveness of the toxicity. For instance, a player highlighted that the competitiveness of game design could induce toxicity:

> Playing this game solo/duo[14] is not fun in the long run when trying to climb ranks. I only play this game now with friends and never read chat, if someone spam click pings, I mute them.

[13]"Ranked game" refers to a game mode where players compete with each other to earn League Points and to increase their standing on a ranked ladder.

[14]"Solo/Duo Queue" means a gameplay option where a player is able to queue by themselves or with the help of another player to move up to higher rankings in LoL's ranking system.

Table 3: Standardized Co-occurrence Rates (decimal values) across 266 threads, comparing combined code frequencies between RQ1 and RQ2 datasets. Higher values indicate more frequent co-occurrences, while values such as -7.484 for RQ1-AttriToxic and RQ2-MorEviEnqu indicate a notably infrequent pairing due to initial non-occurrence value. Refer to Section 4.3 for the rate's definition and formula. Frequencies for RQ1 themes (left) and RQ2 themes (right) are provided in parentheses. The table rows represent eight RQ1 themes, such as Attribution to Toxicity and Introspection, while columns cover eight RQ2 themes, including Future Behavior Suggestion and Support for Moderation Decisions.

RQ1	RQ2							
	RQ2-AttriToxic	RQ2-CompDesig	RQ2-DisagreMem	RQ2-FubeSug	RQ2-InfoProPuni	RQ2-MorEviEnqu	RQ2-ResoModExp	RQ2-SupMod
RQ1-AttriToxic	0.574 (35, 100)	-0.120 (35, 97)	-0.425 (36, 84)	0.336 (35, 379)	0.293 (35, 339)	-7.484 (35, 55)	-0.048 (35, 201)	0.211 (35, 253)
RQ1-CompDesig	0.178 (91, 100)	0.185 (91, 97)	0.061 (91, 84)	0.417 (91, 379)	0.408 (91, 339)	-0.304 (91, 55)	0.082 (91, 201)	0.442 (91, 253)
RQ1-DetailContext	0.269 (91, 100)	0.078 (174, 97)	0.088 (174, 84)	0.615 (174, 379)	0.556 (174, 339)	0.241 (174, 55)	0.297 (174, 201)	0.504 (174, 253)
RQ1-DisagreMem	-0.244 (25, 100)	-0.917 (25, 97)	-0.091 (25, 84)	0.309 (25, 379)	0.271 (25, 339)	-0.238 (25, 55)	-0.415 (25, 201)	0.534 (25, 253)
RQ1-InfoSeek	-0.001 (114, 100)	-0.210 (114, 97)	-0.175 (114, 84)	0.387 (114, 379)	0.510 (114, 379)	-0.469 (114, 55)	0.112 (114, 201)	0.056 (114, 253)
RQ1-Introsp	0.158 (88, 100)	0.070 (88, 97)	0.069 (88, 84)	0.627 (88, 379)	0.347 (88, 339)	-0.207 (88, 55)	0.265 (88, 201)	0.408 (88, 253)
RQ1-NegEmotion	-0.304 (18, 100)	-0.046 (18, 97)	-0.461 (18, 84)	0.248 (18, 379)	0.218 (18, 339)	-0.697 (18, 55)	-0.334 (18, 201)	0.059 (18, 253)
RQ1-PerceUnfair	0.270 (151, 100)	0.061 (151, 97)	0.217 (151, 84)	0.588 (151, 379)	0.539 (151, 339)	0.200 (151, 55)	0.274 (151, 201)	0.558 (151, 253)

5.2.8 Disagreement with Community Members. This theme captures instances of disagreement among players about shared moderation experiences within the LOL community. For instance, fellow players challenged penalized players' justifications for their past behaviors. An illustrative comment was, "*Sounds like you're just making excuses to be mean to people over a video game.*"

Furthermore, players expressed skepticism towards comments from those who were not the original poster but shared their experiences with moderation. For instance, one player expressed doubt by stating, "*You 100% believe OP (poster) that they got a chat restriction from this, and it happened to you. But OP got no chat restriction. Sounds fishy.*" This comment demonstrates a player's skepticism about the credibility of another player's account.

5.3 RQ3: Communication between Punished Players and Peers – Peer Support's Potential Benefits

As shown in Table 3, the co-occurrence analysis highlights that when punished players LOL discussed their actions, a mutual recognition of the game's toxic culture is evident. Specifically, in instances where punished players attributed their actions to the game's toxicity (frequency = 35), peer support often concurred with this viewpoint (frequency = 100), as indicated by a standardized co-occurrence rate of 0.574.

In addition, our co-occurrence analysis suggests three potential benefits of peer support. First, peer support could be instructive in guiding moderated players towards self-improvement and positive behavior changes. Conversations initiated by moderated players on topics like introspection (frequency = 88) or perceived unfairness (frequency = 151) commonly elicited constructive feedback with standardized co-occurrence rates between 0.248 and 0.627, suggesting a tendency to provide helpful advice (i.e., all cells in RQ2-FuBeSug column).

Second, prior work has reported that punished users request platforms' accountability for offering more information about moderation (e.g., [95, 97]). We found that in LoL, peer support often involves sharing insights into the moderation process. When moderated players sought information about their punishments (frequency = 114) or the context (frequency = 174), the response from peers frequently involved informative exchanges, with co-occurrence rates from 0.218 to 0.556, reflecting the informative nature of peer support. (i.e., all cells in RQ2-InfoProPuni column).

Last, while aspects of informativeness and instructiveness were prominent, peer support also encompasses critical dimensions. Our analysis revealed that when punished players discussed topics like the perceived unfairness of punishments (frequency = 151) or specific details of the moderation context (frequency = 174), fellow players also tended to respond with critiques or even sarcasm to endorse the legitimacy of moderation decisions, as reflected by co-occurrence rates around 0.5 (i.e., many cells in RQ2-SupMod). These findings underscore the multifaceted nature of peer support regarding moderation within the LoL.

6 DISCUSSION

In this study, we reported an analysis of how punished players shared their moderation experiences and how community members

responded. Drawing from the peer support literature (e.g., [65]), we conceptualized the experiential sharing as requesting peer support, and community response as providing peer support. The peer support lens helped reveal what kinds of support moderated users need, and the important role that the online community has played in providing such support.

When moderation ends at the point of effecting moderation decisions and impacting target players, the community steps in to address unmet player needs. What we observed in this study about the limits of moderation and the potential role of community provides ample insights into rethinking and designing community-based rehabilitation in the context of video game moderation.

6.1 Contextualizing Peer Support in Online Game Moderation

Punished users constitute an important stakeholder group in online moderation. They are directly impacted by moderation decisions, and are expected to take actions, either improving their future behavior or leaving the platform. Thus, moderation researchers have been interested in understanding how users experience moderation decisions, especially what they do after moderation (e.g., [22, 40, 79]). Extending this literature, we provided a detailed account of how LoL players affected by moderation decisions worked with the rest of the community to make ends meet. While prior moderation research discussed how moderated users question the legitimacy of moderation and devise strategies to bypass moderation [8, 35], such as using language variants [34] or turning off in-game chat function to avoid being toxic in competitive games [52], what we found in this study is a diverse range of player reactions to moderation decisions: Some punished players questioned the fairness and transparency of moderation decision-making, others explored explanations that lead to their own toxic deeds, and still, others accepted the verdicts on them and sought ways to improve future behavior. In return, the community responses also included diverse types of support.

Our findings about peer support highlight the uniqueness and complexity of combatting toxicity in the online gaming context. Punished players tend to desire more information and transparency for their punishments in terms of both notification and explanation of their penalties [70], as such information could help them engage in problem-focused coping [15]—viewing their own behaviors that are punished as a problem and seeking information to solve this problem. Punished players' informational needs also manifest in how players described their punishments as we reported in RQ1: it is not always straightforward to players as to why they were considered toxic and subsequently why they were the sole party held responsible for the in-game toxicity. Instead, they would cite various factors such as the emotional dimension, perceived issues with moderation, the gameplay context, and the larger gamer culture that co-contribute to their in-game toxicity. This highlights the multi-faceted, contextual nature of toxicity. Compared to text-based toxicity, in-game toxicity is exponentially more difficult to detect and adjudicate on. For example, a LoL player could be secretly sabotaging their teamwork without being noticed by teammates [50]. In other words, there are no simple, operational principles for players to readily determine whether a behavior is toxic or not. Hence, when punished players seek peer support, informational support, one of the four primary types of peer support [39], becomes primary as it is highly useful in supplying them normative knowledge to deepen their understanding of their past behavior.

6.2 The Rehabilitative Value of Peer Support

Our study reflects two gaps between moderation and community in the retributive model of moderation. First is the epistemic gap, where the moderation system and the community maintain different knowledge about what constitutes toxic behavior. Our findings showed that punished players perceived the unfairness of moderation because toxic behaviors in more complex forms remained unpunished, such as the 'Vi' example of intentionally ruining others' experiences by taking away their resources in game. Such sentiment was echoed by the responses from other players. In other words, the behaviors the moderation system targeted were not the same as the behaviors the community perceived as toxic. The epistemic gap reflects that the way LoL's moderation apparatus categorizes certain acts as toxic while others as not could violate players' sense of fairness and trust in the moderation system. Chang and Danescu-Niculescu-Mizil [18] found that Wikipedia users who perceived unfairness in penalties were more likely to reoffend. Reoffending happens when users question the fairness and legitimacy of moderation, and the epistemic gap in LoL risks causing reoffending as punished players do not believe that the operations of moderation are based on a legitimate set of knowledge about player behavior.

Second is the participatory gap, where the moderation system assumes little to no participation of the broader community versus the de facto participation of the community. By the very definition of moderation to facilitate community members' cooperative behavior, the community already plays an active role in alignment with the goals of moderation, such as interpreting norms and policies to moderated users, prescribing future behaviors to avoid moderation, and helping moderated users' behavioral improvement in general. If moderation is the core work, what the community is already doing could be considered as a form of "articulation work" [85] that appears secondary to the core work but is indispensable in supporting the core work. Without such articulation work from the community, an affected user would face more challenges in understanding rules and improving behavior, and naturally commit more toxic behavior. However, articulation work also tends to be invisible and underrecognized [85]. Thus, our work also seeks to give visibility to this form of articulation work that the LoL community has been engaging in to support the core goal of moderation.

While the moderation system performs its retributive justice as designed, widening the epistemic and participatory gaps between moderation and community, the community can potentially meet several rehabilitative goals that have been prescribed by criminological research (e.g., [24, 25, 73]). Importantly, the idea of rehabilitation [75] emphasizes the outcome, or whether the convicted offenders have eventually attained self-growth in skills, social functioning, access to resources, etc., over the process, or what kinds of support resources they can gain to achieve a desired outcome. This helps us contemplate how the characteristics of peer support may uphold the rehabilitation of punished LoL players.

First, peer support offers informational support to moderated users. As we discussed previously, given the highly contextual nature of toxic behavior in the online gaming context, it is reasonably challenging for players, and especially new players, to draw clear lines between toxic and non-toxic behaviors. Thus, the information provided through peer support can be valuable for punished players to reconstruct a normative context in order to understand if and why they were toxic.

Second, our study showed that wrongly convicted cases are rare but do exist. For wrongly convicted users, peer support manifests as a form of appraisal support [39], where fellow players agreed that the moderation system had made a questionable decision and affirmed the punished players. This is where players showed empathy and suggested actionable items such as contacting Riot's support team. This corresponds with how criminologists recognize the importance of promoting offenders' self-determination and personal growth [25]. However, emotional and appraisal support are rare, and much of the peer support was geared towards information seeking and provision.

Third, peer support requests and provisions create a healing space in which both offenders and victims could be involved in conversations. This can foster social acceptance of offenders [83], which leads to their eventual social reintegration into the community. We observed many conversations where the respondents shared their own experiences as victims of toxicity when playing the game. In addition, the act of disclosing one's punishments is no simple matter. Such disclosure of negative events could incur social risks such as criticism and disapproval [4]. Thus, the very fact that these punished players were willing to perform such disclosure and initiate conversations with their peers indicate their active pursuit of healing through connecting with other people in the community, openness to potential negative feedback, and, sometimes, desire for self-improvement. In this regard, even if the peer support players received is largely unempathetic information and behavior suggestions, instead of emotional support, it is still of a supportive nature and satisfies informational needs of punished players.

Taken together, we view peer support as a strong mechanism that exists in a naturalistic community setting that is dialogical, interactive, and dynamic. It is applaudable that punished players and others engage in conversations around a hard topic, through which mutual understanding could emerge and toxic players could start to internalize commonly held norms. Even when informational support constitutes a majority of peer support, it is conditioned in an online gaming context, and an essential step towards rehabilitative outcomes.

Seering [87] contrasts two perspectives in moderation research, one focused on platforms and policies-focused perspective that tends to be centralized and top-down, and the other focused on community self-moderation that is decentralized and bottom-up, and advocates for the greater role of community in future online moderation. In a similar vein, moderation researchers have explored various ways to engage the community in moderation, such as creator-led moderation [44], co-designing moderation with marginalized people [96], or rethinking moderation adjudication by involving user voice [77]. Our study resonates with this direction in emphasizing the potential of community in the provision of peer support during the post-moderation phase, even when the moderation is done in a platforms and policies-driven manner. Yet, different from seeing a clear distinction between those two perspectives, what our study identified is how two perspectives work together: platforms and policies-driven moderation functions as a primary mode of managing player behavior, but community plays a complementary role in facilitating punished players' behavior improvement.

Peer support, here, should not be conflated with the appeal process that is usually offered by large-scale, commercial platforms that employ a sophisticated combination of automated and human moderation [80, 81]. The appeal process is instituted by platforms and functions more or less as a bureaucracy [69], sometimes widens inequality between users [2], and do not always address players' concerns [70], while peer support is community-led, dynamic, and contextual. In this context, although peer support providers are appointed moderators, they de facto participate in the process of community management, and thus take on some of the social roles that Seering et al. identified among volunteer moderators [88], such as nurturing and supporting communities.

While recent moderation literature has started to suggest online platforms involve user education in moderation to help punished users better learn about content rules and reform behaviors (e.g., [97]), what we observed in this study is punished players' social learning practices. The former stresses a top-down approach to disseminate essential knowledge to people, more or less mimicking classroom-based learning, the latter stresses a bottom-up approach to the acquisition of knowledge, emphasizing the social and ubiquitous nature of learning [60]. The distinction is important because the player community possesses situated knowledge that is not captured in automated moderation, and because punished players are actively seeking answers, with or without support from the platform. As shown in our study, punished players interacted with fellow users to make sense of their moderation experiences, and sought more information about their behavior, the community's common understanding, and the moderation system. This is the type of learning that happens in the wild and not through education.

6.3 Implications for Designing Community-Based Rehabilitation Mechanisms to Support Moderated Users

Our study revealed punished players' needs for community-based rehabilitation (CBR), and opportunities for community members to engage. Specifically, we outline five player needs with design considerations for CBR:

Communicative Need: Punished players have the need to start a conversation, sharing their moderation experiences and seeking input from the community. Having a conversation with someone else about the moderation experience can become the first step towards rehabilitation.

Informational Need: Punished players desire to understand how moderation systems make decisions in order to improve future behavior or avoid future moderation punishment. Punished players' informational need has been a focus of moderation research. For example, Jhaver et al. tested whether explanations for content removal could help moderated users to learn [43]. Similar to how users collectively generate folk knowledge of moderation decisions

[67], our findings further point out that peers can specify or highlight the extent of credibility of their knowledge or information to relieve punished players' concerns and questions. And such extent of credibility, as our findings showed, ranges from platform rule references to guesswork on reasons why moderation happens.

Emotional Need: Punished players usually express negative emotions such as frustration and sadness when describing a moderation decision. Thus, they also seek emotional support from their fellow community members, and sometimes, the community might respond with empathy or sharing similar moderation experiences. However, not every response to such emotional needs is positive since peer support can be critical or even sarcastic. Then a design consideration surfaces: Online platforms could take more accountability in peer support communication by visualizing the sentiment of peer support to both shape the communication to be informative and prevent new toxicity from emerging in such communication. Especially, as Section 5.3 showed, future behavior suggestions and information provision about moderation are two conspicuous utilities of peer support compared to other types of it, indicating that informing the measured sentiment of peer support can allow punished players to better locate and identify informative and instructive feedback.

Need for Moderation Fairness: Punished players desire to be treated fairly by the moderation system. Upon perceiving uneven treatment across players, they hope that procedural fairness can be restored in meaningful ways. Related to this, Vaccaro et al. explored ways to support users in contesting moderation decisions [96]. But it is worth noting that expressing the perceived unfairness of moderation should not be the equivalent of sharing negative emotions only. We found that the community can respond to punished players by asking for more evidence of not deserving punishments or merely disagreeing with punished players' perceived unfairness. Thus, when punished players express their deemed unfairness, there should be design elements reminding them of how negative the sentiment of their language use is, whether they have evidence to share (e.g., screenshots), etc., to ensure further communication could be evidence-driven and objective.

Need for Reflection: A moderation decision could trigger a "teachable moment" [72] for punished players to reflect on their past behavior and contemplate behavior change. Our findings showed how punished players performed introspection as to how they committed toxicity. Reflection is a necessary pathway to better behavior and aligns strongly with the very goal of moderation but has not received much design consideration. A peer support mechanism can encourage punished players to initiate introspections. Thus, whatever platforms affording communication among players could consider stressing the meaningfulness of community support for reforming problematic behaviors and shaping productive or playful community members before users respond peer support.

Besides the abovementioned core player needs and design considerations, our study has further implicated how a holistic peer support mechanism could be designed as post-moderation support. While an online forum, as analyzed in this study, could support peer support, it is not designed for this purpose. As such, our analysis also revealed several shortcomings of using an online forum. For example, peer support requests could be answered unevenly, where some threads received numerous replies while some received only a few. Punished players also need guidance on how to write a peer support request, such as what kinds of essential information should be included. In a similar vein, Support providers could also be better guided in terms of how to provide helpful support. Thus, we envision that punished players could be provided with a system that involves punished players, others who do not experience moderation, and representatives from online platforms to engage in communication together. Especially, this system can allow punished players to enter a description of their moderation experience, and community members or platform officials are allowed to review such description and provide helpful feedback.

6.4 Limitations and Future Work

Our study does not generalize to describe the whole picture of peer support and moderation experience sharing in LoL. However, the combination of content analysis, the thoughtful co-occurrence formula generated to avoid confounding factors, and random sampling to preserve the representativeness of original data [47] represent a systematic, reproducible workflow that helped us obtain an initial understanding of peer support communication among players. And we expect future moderation research to generalize the themes we found and implement them in designing peer support mechanisms for more transparent and fairer moderation, as more researchers called for (e.g., [30, 95]), benefiting those who offer evidence and believe they are unfairly moderated. Future research can build on this study to conduct interviews or survey with punished players to understand their experiences with seeking and receiving support from community members. Games researchers can also follow a similar methodology to examine other online game communities to cross-validate findings from this study.

7 CONCLUSION

In this study, we analyzed peer support requests and provisions for punished players in the online game community of League of Legends. Our findings characterized peer support mechanisms that could support moderated users during the post-moderation phase, which current moderation systems rarely pay attention to, but the community plays an indispensable role in. Informed by insights from criminology, we see post-moderation support as an important way to enhance existing moderation systems via bridging the gap between moderation and community and adding rehabilitative values. We call for more attention to documenting and conceptualizing existing community resources that support punished offenders, as well as design efforts to integrate rehabilitative values into moderation systems.

ACKNOWLEDGMENTS

Many thanks to the reviewers for their thoughtful feedback. The work is partly supported by the National Science Foundation (#2326505 and #2334934).

REFERENCES

[1] 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*, 2018. ACM Press, 365–372. . https://doi.org/10.1145/3270316.3271545

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

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14, (May 2020), 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

[4] Natalya N. Bazarova and Yoon Hyung Choi. 2014. Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites. *Journal of Communication* 64, 4 (August 2014), 635–657. https://doi.org/10.1111/jcom.12106

[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). https://doi.org/10.1145/3411764.3445157

[6] James Bonta and D. A. Andrews. 2016. *The Psychology of Criminal Conduct*. Taylor & Francis.

[7] Charles M. Brooks and Janice L. Ammons. 2010. Free Riding in Group Projects and the Effects of Timing, Frequency, and Specificity of Criteria in Peer Assessments. http://dx.doi.org/10.1080/08832320309598613 78, 5 (May 2010), 268–272. https://doi.org/10.1080/08832320309598613

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

[9] Stéphane Brutus and Magda B. L. Donia. 2017. Improving the Effectiveness of Students in Groups With a Centralized Peer Evaluation System. https://doi.org/10.5465/amle.9.4.zqr652 9, 4 (November 2017), 652–662. https://doi.org/10.5465/AMLE.9.4.ZQR652

[10] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, April 2008. ACM Press, 1101–1110. . https://doi.org/10.1145/1357054.1357227

[11] Colin Campbell. 2014. How Riot Games encourages sportsmanship in League of Legends. *Polygon*. Retrieved August 27, 2023 from https://www.polygon.com/2014/3/20/5529784/how-riot-games-encourages-sportsmanship-in-league-of-legends

[12] Alessandro Canossa, Dmitry Salimov, Ahmad Azadvar, Casper Harteveld, and Georgios Yannakakis. 2021. For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay. *Proceedings of the ACM on Human-Computer Interaction* 5, CHIPLAY (October 2021). https://doi.org/10.1145/3474680

[13] Robyn Caplan and Tarleton Gillespie. 2020. Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society* 6, 2 (April 2020), 205630512093663. https://doi.org/10.1177/2056305120936636

[14] Kevin M. Carlsmith, John M. Darley, and Paul H. Robinson. 2002. Why do we punish? deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83, 2 (2002), 284–299. https://doi.org/10.1037/0022-3514.83.2.284

[15] Charles S. Carver, Michael F. Scheier, and Jagdish K. Weintraub. 1989. Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology* 56, 2 (1989), 267–283. https://doi.org/10.1037/0022-3514.56.2.267

[16] Fisher Cd. 1986. Organizational socialization: an integrative review. *Res Pers Hum Res Manag* 4, (1986), 101–145.

[17] Christina M. Cestone, Ruth E. Levine, and Derek R. Lane. 2008. Peer assessment and evaluation in team-based learning. *New Directions for Teaching and Learning* 2008, 116 (December 2008), 69–78. https://doi.org/10.1002/TL.334

[18] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *The World Wide Web Conference* (*WWW '19*), May 13, 2019, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 184–195. . https://doi.org/10.1145/3308558.3313638

[19] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How Community Feedback Shapes User Behavior. In *ICWSM*, May 2014. . Retrieved from http://arxiv.org/abs/1405.1429

[20] Chipteck. 2022. In-Client Penalty Notifications FAQ. *League of Legends Support*. Retrieved August 28, 2023 from https://support-leagueoflegends.riotgames.com/hc/en-us/articles/205097293-In-Client-Penalty-Notifications-FAQ

[21] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. 2010. Socialization tactics in wikipedia and their effects. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2010), 107–116. https://doi.org/10.1145/1718918.1718940

[22] Simon Copland. 2020. Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review* 9, 4 (2020), 1–26. https://doi.org/10.14763/2020.4.1516

[23] Francis T. Cullen, Robert Agnew, Harry Allen, Todd Clear, John Eck, David Farrington, Anthony Petrosino, Alex Piquero, Lawrence Sherman, Benjamin Steiner, James Unnever, John Wozniak, and John Paul Wright. 2005. THE TWELVE PEOPLE WHO SAVED REHABILITATION: HOW THE SCIENCE OF CRIMINOLOGY MADE A DIFFERENCE. *Criminology* 43, 1 (February 2005), 1–42. https://doi.org/10.1111/J.0011-1348.2005.00001.X

[24] Francis T. Cullen and Karen E. Gilbert. 2012. *Reaffirming Rehabilitation: Second Edition*.

[25] Andrew Day and Tony Ward. 2010. Offender Rehabilitation as a Value-Laden Process. *International Journal of Offender Therapy and Comparative Criminology* 54, 3 (May 2010). https://doi.org/10.1177/0306624X09338284

[26] Alan R. Dennis and Joseph S. Valacich. 1999. Rethinking media richness: Towards a theory of media synchronicity. *Proceedings of the Hawaii International Conference on System Sciences* (1999), 12. https://doi.org/10.1109/HICSS.1999.772701

[27] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit. In *Conference on Human Factors in Computing Systems - Proceedings*, May 2019. Association for Computing Machinery. . https://doi.org/10.1145/3290605.3300372

[28] Jordan Eschler, Zakariya Dehlawi, and Wanda Pratt. 2015. Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (2015), 101–109. https://doi.org/10.1609/icwsm.v9i1.14611

[29] esports.com. 2021. Bans Up to 14 Days - Riot introduces harsher measures against AFKs and Dodging in LoL. *esports.com*. Retrieved August 27, 2023 from https://www.esports.com/en/bans-up-to-14-days-riot-introduces-harsher-measures-against-afks-and-dodging-in-lol-206777

[30] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020. Association for Computing Machinery (ACM), 1–14. . https://doi.org/10.1145/3313831.3376293

[31] Rosta Farzan, Robert Kraut, Aditya Pal, and Joseph Konstan. 2012. Socializing volunteers in an online community: A field experiment. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2012), 325–334. https://doi.org/10.1145/2145204.2145256

[32] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–28. https://doi.org/10.1145/3392845

[33] David Garland. 2012. The Culture of Control: Crime and Social Order in Contemporary Society - David Garland - Google Books.

[34] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (December 2018), 4492–4511. https://doi.org/10.1177/1461444818776611

[35] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* 5, 1 (January 2019), 205630511983258. https://doi.org/10.1177/2056305119832588

[36] Tarleton Gillespie. 2018. Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

[37] Eric Goldman. 2021. Content Moderation Remedies. *Michigan Technology Law Review* (2021). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id$=$3810580

[38] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, (2015).

[39] Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice* 4, (2008), 189–210.

[40] 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. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (November 2019), 1–33. https://doi.org/10.1145/3359294

[41] 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. https://doi.org/10.1145/3338243

[42] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021), 30. https://doi.org/10.1145/3479525

[43] Shagun Jhaver, Amy Buckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 27.

[44] Shagun Jhaver, Quanze Chen, Detlef Knauss, and Amy Xian Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022. ACM Press.

[45] 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. https://doi.org/10.1145/3185593

[46] Cheryl Lero Jonson, Francis T. Cullen, and Jennifer L. Lux. 2013. Creating Ideological Space. In *What Works in Offender Rehabilitation*. John Wiley & Sons, Ltd, 50–68. https://doi.org/10.1002/9781118320655.ch3

[47] Hwalbin Kim, S. Mo Jang, Sei-Hill Kim, and Anan Wan. 2018. Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society* 4,. https://doi.org/10.1177/2056305118772836

[48] John Kleinig. 2008. *Ethics and Criminal Justice: An Introduction*. Cambridge University Press.

[49] Jason Koebler and Joseph Cox. 2018. Content Moderator Sues Facebook, Says Job Gave Her PTSD. *Vice*. Retrieved June 23, 2023 from https://www.vice.com/en/article/zm5mw5/facebook-content-moderation-lawsuit-ptsd

[50] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the 2020 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '20*, November 2020. Association for Computing Machinery (ACM), 81–92. . https://doi.org/10.1145/3410404.3414243

[51] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021). https://doi.org/10.1145/3476075

[52] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (October 2020), 1–27. https://doi.org/10.1145/3415173

[53] Yubo Kou and Xinning Gui. 2020. Emotion Regulation in eSports Gaming: A Qualitative Study of League of Legends. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (October 2020), 1–25. https://doi.org/10.1145/3415229

[54] 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 ACM SIGCHI Conference on Human factors in computing systems*, 2021. .

[55] Yubo Kou and Bonnie Nardi. 2014. Governance in League of Legends: A Hybrid System. In *Foundations of Digital Games*, 2014. .

[56] Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology*. SAGE Publications.

[57] Klaus Krippendorff. 2004. Reliability in Content Analysis. *Human Communication Research* 30, 3 (2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[58] Haewoon Kwak and Jeremy Blackburn. 2014. Linguistic Analysis of Toxic Behavior in an Online Video Game. In *International Conference on Social Informatics*, 2014. .

[59] Greg Lastowka. 2011. *Virtual Justice: The New Laws of Online Worlds*. Yale University Press. Retrieved from http://www.amazon.com/dp/0300177747

[60] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation* (1st ed.). Cambridge University Press.

[61] Sung Je Lee, Eui Jun Jeong, and Joon Hyun Jeon. 2019. Disruptive behaviors in online games: Effects of moral positioning, competitive motivation, and aggression in "league of legends." *Social Behavior and Personality* 47, 2 (2019). https://doi.org/10.2224/SBP.7570

[62] Nutan Lele. 2022. How Many People Are Playing League of Legends in 2022? *AFK Gaming*. Retrieved August 27, 2023 from https://afkgaming.com/esports/how-many-people-are-playing-league-of-legends-in-2022

[63] Jeffrey Lin. 2013. The Science Behind Shaping Player Behavior in Online Games. In *Game Developers Conference*, 2013. .

[64] Mark W. Lipsey and Francis T. Cullen. 2007. The Effectiveness of Correctional Rehabilitation: A Review of Systematic Reviews. *Annual Review of Law and Social Science* 3, 1 (2007), 297–320. https://doi.org/10.1146/annurev.lawsocsci.3.081806.112833

[65] Ngar-Fun Liu and David Carless. 2006. Peer feedback: the learning element of peer assessment. *Teaching in Higher Education* 11, 3 (July 2006), 279–290. https://doi.org/10.1080/13562510600680582

[66] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–25. https://doi.org/10.1145/3449180

[67] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* (2021). 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. *Proceedings of the ACM on Human-Computer Interaction* (2022). 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 *CSCW'22 Companion: Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, November 2022. Association for Computing Machinery (ACM), 8–13. . https://doi.org/10.1145/3500868.3559445

[70] Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. . https://doi.org/10.1145/3544548.3581097

[71] Abhishek Mallick. 2022. League of Legends patch 12.10 to reward players with special recalls if they reach Challenger or Honor level 5. Retrieved August 27, 2023 from https://www.sportskeeda.com/esports/news-league-legends-patch-12-10-reward-players-special-recalls-reach-challenger-honor-level-5

[72] C. M. McBride, K. M. Emmons, and I. M. Lipkus. 2003. Understanding the potential of teachable moments: the case of smoking cessation. *Health Education Research* 18, 2 (April 2003), 156–170. https://doi.org/10.1093/her/18.2.156

[73] Fergus McNeill. 2012. Four forms of 'offender' rehabilitation: Towards an interdisciplinary perspective. *Legal and Criminological Psychology* 17, 1 (February 2012), 18–36. https://doi.org/10.1111/J.2044-8333.2011.02039.X

[74] Subhradeep "Bucketbaba" Mukherjee. 2021. Is Riot's AFK penalty system a bit too unfair and severe for Indian Valorant players? Retrieved August 27, 2023 from https://www.sportskeeda.com/valorant/is-riot-s-afk-penalty-system-bit-unfair-severe-indian-valorant-players

[75] National Institutute of Justice. 2020. Practice Profile: Rehabilitation Programs for Adult Offenders. *CrimeSolutions, National Institute of Justice*. Retrieved June 24, 2023 from https://crimesolutions.ojp.gov/ratedpractices/101

[76] Jamie Newsome and Francis T. Cullen. 2017. The Risk-Need-Responsivity Model Revisited: Using Biosocial Criminology to Enhance Offender Rehabilitation. *Criminal Justice and Behavior* 44, 8 . https://doi.org/10.1177/0093854817715289

[77] 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. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (April 2022). https://doi.org/10.1145/3512929

[78] Alex R. Piquero, Francis T. Cullen, James D. Unnever, Nicole L. Piquero, and Jill A. Gordon. 2010. Never too late: Public optimism about juvenile rehabilitation. *Punishment & Society* 12, 2 (August 2010). https://doi.org/10.1177/1462474509357379

[79] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021). https://doi.org/10.1145/3476057

[80] Sarah T. Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications* (January 2016).

[81] Sarah T. Roberts. 2019. *Behind the screen: content moderation in the shadows of social media*. Yale University Press.

[82] Natti Ronel. 2006. When good overcomes bad: The impact of volunteers on those they help. *Human Relations* 59, 8 (August 2006), 1133–1153. https://doi.org/10.1177/0018726706068802

[83] Natti Ronel and Ety Elisha. 2010. A Different Perspective: Introducing Positive Criminology. *International Journal of Offender Therapy and Comparative Criminology* 55, 2 (January 2010), 305–325. https://doi.org/10.1177/0306624X09357772

[84] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021), 33. https://doi.org/10.1145/3479512

[85] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)* 1, 1–2 (March 1992), 7–40. https://doi.org/10.1007/BF00752449

[86] Sarita Schoenebeck, Oliver L. Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society* (March 2020), 146144482091312. https://doi.org/10.1177/1461444820913122

[87] Joseph Seering. 2020. Reconsidering Community Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (October 2020), 28. https://doi.org/10.1145/3415178

[88] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (March 2022), 621–640. https://doi.org/10.1177/1461444820964968

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

[90] Richard J. Siegert, Tony Ward, William M. M. Levack, and Kathryn M. Mcpherson. 2007. A Good Lives Model of clinical and community rehabilitation. *Disability and Rehabilitation* 29, 20–21 (January 2007), 1604–1615. https://doi.org/10.1080/09638280701618794

[91] Spandana Singh. 2019. Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. New America.

[92] Todd Spangler. 2018. ESPN+ to Live-Stream 'League of Legends' Esports Events Under New, Non-Exclusive Pact. *Variety*. Retrieved August 27, 2023 from https://variety.com/2018/digital/news/espn-league-of-legends-esports-disney-1202821402/

[93] Nicolas P. Suzor. 2010. The role of the rule of law in virtual communities. *Berkeley Technology Law Journal* 25, 4 (2010), 1818–1886.

[94] Katie Salen Tekinbaş, Krithika Jagannath, Ulrik Lyngs, and Petr Slovák. 2021. Designing for Youth-Centered Moderation and Community Governance in Minecraft. *ACM Trans. Comput.-Hum. Interact.* 28, 4 (July 2021), 24:1-24:41.

https://doi.org/10.1145/3450290

[95] 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. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (October 2020). https://doi.org/10.1145/3415238

[96] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021), 28. https://doi.org/10.1145/3476059

[97] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media and Society* 20, 11 (2018), 4366–4383. https://doi.org/10.1177/1461444818773059

[98] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Conference on Human Factors in Computing Systems - Proceedings*, May 2019. Association for Computing Machinery. . https://doi.org/10.1145/3290605.3300390

[99] Bingjie Yu, Katta Spiel, and Leon Watts. 2020. Supporting Care as a Layer of Concern: Nurturing Attitudes in Online Community Moderation. *Proceedings of CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3334480.3383009