Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community

YUBO KOU, Pennsylvania State University, USA

Platforms face the challenge of managing toxic behaviors such as flaming, hateful remarks, and harassment. To discipline their users, platforms usually adopt a punitive approach that issues punishments ranging from a warning message to content removal to permanent ban (PB). As the severest punishment, PB deprives the user of their privileges on the platform, such as account access and purchased content. But little is known regarding the experiential side of PB within the user community. In this study, we analyzed PB in League of Legends, one of the largest online games today. We argue that what PB does is not precisely to discipline players into well-behaved community members. Rather, PB functions to produce the stereotype of "the most toxic player" in the community and is best seen as a platform rhetoric. We further discuss the need to contextualize toxicity from the restorative lens.

CCS Concepts: • **Human-centered computing** ~ Human computer interaction (HCI); •Human-centered computing ~ Collaborative and social computing

KEYWORDS: Permanent ban, punishment, toxicity, governance, community moderation, League of Legends, player community, discursive field, discourse analysis

ACM Reference format:
Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. In *Proceedings of the ACM on Human-Computer Interaction*, Vol 5, No. CSCW2, Article 334 (October 2021). 21 pages. https://doi.org/10.1145/3476075

1 INTRODUCTION

Moderation describes the practice online platforms engage in to curb rampant disruptive behaviors such as flaming, harassment, and hate speech. Moderation includes a series of coordinated mechanisms such as platform policies and technical procedures to enforce them such as flagging, aggregation of flags, adjudication, and punishment [31]. Moderation often borrows from legal practices [70], and mimics a legal process composed of several steps: First, a user's behavior is flagged, and a moderation system adjudicate on the flagged behavior. The user, once convicted by the moderation system, receives punishment of various forms ranging from post removal [39] to permanent ban (PB) [48]. Coincidentally, the recent wave of account suspensions of the U.S. President Donald Trump from multiple major social media sites [14] highlights the contentious nature of account suspension.

While recent years have seen a surging body of research on moderation, examining moderation tools (e.g., [38,40]), and moderator experiences (e.g., [17,56,81]), punishment as the last step of moderation has rarely been at the analytic focus of the moderation literature, with a few exceptions (e.g., [37,39]). Punishment, however, is no simple matter. It is a retributive act that seeks to inflict deprivation on the person being punished [3]. Punishment on online

This work is supported by the National Science Foundation, under grant IIS-2006854.
Author's address: Yubo Kou, College of Information Sciences and Technology, Penn State, PA 16801 USA. yubokou@psu.edu

platforms manifests as a form of deprivation of privileges to punished users, and a threat of deprivation to others. PB is usually the universal and ultimate form of punishment that platforms carry out against their users. Although the goal of punishment is often proclaimed to discipline, so that individuals can internalize and behave to the normative expectations of community norms and platform rules, little is known regarding the actual experiential implications of PB. In addition, the criminal justice literature suggests that even with justified conviction and sentence, the same punishment could inflict deprivations of different severities and experiences among different people [43].

Existing challenges and controversies related to online moderation amount to the urgent need of investigating how users themselves experience PB and what is the disciplinary effect of PB and reflecting on whether and how PB should remain as the ultimate punishment in moderation, as well as who can make such moderation decisions. These important questions motivate us to undertake a project unpacking users' subjective experience with PB. Specifically in this paper, we present an empirical study of PB in League of Legends (LoL), one of the most popular online games. The game uses PB against toxic players it deems to have severely violated its terms of service and code of conduct. Toxic behavior refers broadly to user behavior that is "socially unacceptable and disrupt other people's online experience" [49]. Common types of toxic behaviors include harassment, racial slur, personal attack, flaming, and hate speech. (In the rest of the paper, we use toxic and disruptive interchangeably.) We collected 197 player discussion threads from the '/r/leagueoflegends' subreddit, one of the largest LoL-specific online forums with more than five million subscribers.

We draw from the notions of discourse and discursive field to frame LoL players' experiences with PB. Discourse is "made up of a limited number of statements for which a group of conditions of existence can be defined" [21]. Discourse is a relatively stable system of language that determines possible speech acts by people in a discursive field [21]. A discursive field consists of "competing ways of giving meaning to the world and of organizing social institutions and processes" [79]. When players make concrete speech acts, such as expressing support for or opposition against PB, their speech acts are governed by discourses.

Through a combination of thematic analysis [6] and discourse analysis [26], we identified five distinct player discourses centered on PB: PB as retribution, as community purification, as platform economy, as recirculation, and as black box. These different discourses converge at associating PB with the "most toxic players," but diverge at PB's nature and effects. These varying and even contradicting framings of PB reflects the complexity of discipline and punishment at the intersection of community moderation and video gamers' culture of toxic meritocracy [61]. PB remains relevant to player experience, but its intended role as a disciplinary device is deconstructed and mythicized within the player community. PB regresses towards a platform rhetoric. The existing use of PB reflects a retributive framing of toxicity that unifies toxic behavior, account, and player, which should be viewed as distinct constructs with overlapping meanings. This analysis of PB, the severest form of punishment, points to a larger question of how to redesign the current punitive logic of moderation systems, which aligns with the growing attention to restorative and participatory values within HCI and CSCW [2,10,68,69].

The present study makes several contributions to the HCI and CSCW literature including: 1) empirical findings can bring detailed documentation of moderation and provoke new ways of moderation design; 2) conceptual insights could deepen understanding of effective and sustainable game moderation (sustainable is construed in consideration of toxic meritocracy); 3)

broadened understandings of online moderation could bring more scholarly attention to punishment in existing moderation literature, especially since punishment is often overlooked in the past.

2 RELATED WORK

We first draw from Michel Foucault's work on discipline and punish to discuss the core terms and theoretical considerations to be referenced in the rest of the paper. We then review the existing moderation literature with a focus on existing discussions of punishment. Lastly, we review existing literature on toxic behaviors in online games.

2.1 Discipline and Punish: From Prison to Platform

Michel Foucault's ideas about discipline, punishment, and power in his book *Discipline and punish: The birth of the prison* [22] provide primary theoretical underpinning for this analysis. Prior to the eighteenth century, punishments were orchestrated spectacles, involving torture or bodily dismemberment in front of the public, so as to induce fear and terror in subjects and manifest the power of the sovereign.

Prison was later reimagined as a form of punishment because of the origin of the idea of *discipline*. Discipline refers to the whole of social and technological means that have been developed and maintained to monitor and control the body's operations. The disciplinary power exercises three primary techniques including hierarchical observation, normalizing judgment, and examination. In modern prisons, inmates are disciplined through an expansive range of means, such as timetables to regulate their time and sense of time, cellular architectures to individuate them, and prison guards' observation and instructions. The imprisonment is rationally calculated based on type and severity of crime. The purpose of prison is not to punish them physically, but to shape their minds, reproducing them as obedient subjects, or "docile bodies." In so doing, the prisoners are considered *reformed* and *normalized*. Not all prisoners can be reformed. Those who can are considered *delinquents*, a category that the prison is to produce in order to justify its existence. Actions of discipline, surveillance, and analyses and experiments of individuals lead to *knowledge* about people, or the modern *human sciences*.

Prison is not the only social institution for disciplinary purposes. It is part of the web of power relations permeating society, complemented by other social institutions such as hospitals, schools, factories and military camps. Foucault used Bentham's *panopticon*, a famous type of prison, to exemplify the disciplinary power. The point of panopticon is to put the individuals under constant surveillance so that every individual acquires a permanent awareness of being monitored, regardless of the actuality, and self-discipline. To further highlight the shift to surveillance as the latest mode of discipline, Foucault claimed that "Our society is one not of spectacle, but of surveillance" [22]. The pervasive and totalitarian discipline gives rise to the neoliberal self in modern societies. People are no longer just collectives. Through the process of individualization, they are transformed into individuals who must take care of and be responsible for their own minds and deeds [23].

Foucault's ideas have been among the underpinnings for contemporary theorizations of discipline and punish on platforms like Facebook, Twitter, and Reddit. At the macro-scale, scholars have pondered the entangled state of online platforms and global capital. Terms such as "surveillance capitalism" [83] seek to describe how online platforms function as modern institutions to structure individuals' experiences, monitor and more importantly produce data about their behaviors, (known as datafication,) and commodify such data and associated

analysis and prediction. Behavioral economics are deliberately integrated into platform design to maximize datafication and thus profitability [30,67]. The notion of surveillance capitalism aligns with Foucault's panoptic surveillance well: interconnected platforms such as Google and Facebook are already sufficiently integrated into people's everyday work and life, putting people under constant and seamless surveillance. It is unsurprising if one encounter a relevant ad on Facebook immediately after they have finished a relevant search on Google. Specifically, Zuboff pointed to how platforms have reinvented their users as merely data producers and shifted focus on the human data market [83]. This characterizes the rampant growth and data harvesting of online platforms which have received limited regulations in the past, although there have always been concerns about privacy and surveillance [16,24].

Platforms are not precisely prison-like institutions that execute the power of the sovereign. Rather, they are primarily corporate actors that compete in a marketplace of platforms and seek to maximize their revenue and profit. Discipline still happens invisibly, but the purpose seems to be transforming users into neoliberal individuals. Some argue that social networks seek to reproduce people in favor of the "neoliberal culture of performance" [10]. Others seem how users are transformed into productive data generators, who constantly engage with the web of platforms. Platforms like Facebook could govern individuals' mediated social interactions and social life and discipline individuals into biopolitical producers whose labor in such forms as interactions and content can be captured for profit [70]. Individuals also become increasingly reliant on Facebook as the central place for the accumulation and maintenance of social capital. Regulation of social interaction, as Schwarz continued [70], is inseparable from its exploitation, because certain users' conduct might threaten profits. The main punitive technology is exclusion, in forms like content removal and account deletion. The threat of these punishments is consequential because of Facebook's status as the "central bank of social capital," and the deprivation of social capital and social interaction is similar to the harms of imprisonment [70].

In recent years, due to the increasing scale and severity of toxicity in forms such as harassment [40], hate speech [11], and online bullying [73], more societal and scholarly attention has been paid to moderation. The slow recognition of the significance of moderation could be attributed to several causes. For example, moderation has been primarily an afterthought, as rapidly growing online platforms like Facebook have not considered moderation as one of their primary tasks from the very beginning [28]. In addition, in the U.S., platforms enjoy the immunity to moderate user-generated content under Section 230, a legislation piece signed into law in 1996. This status of immunity has received increasing scrutiny in recent years as people questioning the boundary between moderation and censorship, the essence of free speech, and the suitability of Section 230 for modern-day platforms [13,29]. Therefore, platforms have started to make or announce significant moves towards disciplining their users, so as to render their content compliant with societally accepted standards. For example, the Facebook CEO proposed to use AI to develop more efficient and accurate moderation techniques [84]. However, these moves are best seen as exploratory efforts, given the scale of controversies many moderation decisions have incurred [28,42,65]. Platforms themselves do not yet have an ideal approach to moderation. Specifically, the disciplinary actions that platforms have taken against their toxic users are confined to variations of punishment, such as content removal and account suspension.

In sum, there is a need to understand how platform users experience punishment as a disciplinary act. We invoke Foucault's notion of discipline because it aligns with what modern platforms ostensibly seek to achieve through their moderation apparatuses, to transform their

users into docile bodies who stay compliant with platform rules. Then, we seek to examine the disciplinary effect of PB in LoL players' subjective experiences, to articulate whether and how discipline happens through PB.

2.2 Moderation of Online Games

Social media platforms have been a research focus in moderation research, but video games have not received much attention [41]. On the one hand, many findings from social media moderation research are translatable to the study of video game moderation. For instance, the sociotechnical mechanisms in social media moderation, such as manual content moderation [15] and automated content moderation [38], could find resemblances in specific human-machine configurations in online game moderation. Recent research has paid much attention to moderators' labor, learning, and experience in the systems [17,41,81,82], as well as the emotional challenges they have encountered when reviewing flagged content [17,42,58,64,81]. Human moderators' struggles and practices are echoed by a few game moderation instances involving human participation [19,48,57]. In addition, a nascent body of literature has investigated how users experience punishments. Gerrard analyzed how users developed better understandings of pro-eating disorder content moderation and devised circumvention strategies [27]. Jhaver et al. showed how providing explanations about content removal on Reddit could help punished users to learn norms [39]. Video games also punish their players in various ways such as chat restriction and account suspension.

On the other hand, several characteristics distinguish moderation of online games from that of social media platforms. First, social media platforms are public venues with low barriers to entry. People could easily create a "throwaway" account [52], or develop a bot to post social media messages [74]. Social media are designed in this way to attract and engage a massive user population. In comparison, online games are a "synthetic world" [9] with specific sets of norms and expectations, which present unique barriers to entry. Regarding this, Christopher Paul coined the term "toxic meritocracy" to describe how online game cultures, facilitated by game design and practiced by gamers, tend to celebrate merit and put extreme focus on skill, achievement, and hard work [61]. Such cultural orientation predisposes players to a narrow set of meritocratic values and marginalizes other meaningful ones such as empathy and collaboration.

Second, today's social media platforms are becoming homogeneous in terms of functions, features, and user experience [59], and it has become common to read stories about users migrating from one platform to another[1]. But video games tend to offer a more or less unique experience, which could not be easily replicated by its competitors. Consequently, many players came back even after multi-year real person bans[2][3].

Third, many moderation decisions on social media platforms are inherently controversial, and could easily become high-profile and capture public concern and sensation; and platform owners are forced to respond, and in some cases, rescind the decisions [28]. In comparison, online game companies have enjoyed a high degree of dominance over in platform governance,

[1] https://www.fastcompany.com/90358305/six-months-after-tumblrs-nsfw-ban-these-kink-communities-are-coming-out-on-top
[2] https://blogoflegends.com/2020/07/23/league-of-legends-riot-iwilldominate-ban/
[3] https://www.riftherald.com/culture/2018/1/4/16850598/tyler1-unbanned-lol-reformed

with little interruption from outside the gaming world, so much so that game moderations could be largely analyzed along the lines of player norms and platform policies.

Game companies rely primarily upon platform policies such as End User License Agreements (EULAs), Terms of Service (TOS), and community code of conduct to govern player behavior in a top-down fashion, may or not involving the minimal participation of players [19,34,48]. Platform policies are legal means for game developers to protect their own commercial interests [1,76]. Platform policies grant game developers the right to punish players, since the players must have already "accepted" the rules before playing the games. To punish players, platform rules usually include a constitutive component that is the classification of toxic behaviors. In so doing, game developers could cite these rules to issue forms of punishments they deem necessary to discipline their player community.

However, platform policies are usually static and only occasionally updated, resulting in conflicts between platform definitions and what players believe constitute disruptive behavior. For instance, platforms face challenges clearly defining whether the behavior of exploiting a bug as cheating or not [35]. In addition, classification systems by nature are a reductive approach to complex social phenomena [5], and could have negative consequences such as misrepresenting or marginalizing certain behaviors [4].

Player experiences could rarely be prescribed by platform rules. Players' evolving gameplay and understandings of behavioral standards are dynamic and evolving [75]. Thus, it is important to explore how players themselves make sense of and respond to permanent bans, as a way to reflect upon game developers' existing moderation practices.

3 BACKGROUND

League of Legends, developed by Riot Games at Santa Monica, California, USA, is one of the largest online games today. Its game genre is multiplayer online battle arena (MOBA), featuring a highly competitive gaming culture. LoL is a match-based game. In its most popular gameplay mode, each match takes between two teams of five players who do not know each other. Players are expected to work together and resolve potential conflicts through fast-paced collaboration. Figure 1 is a screenshot of the beginning of a match. LoL players frequently experience immense frustration and anxiety [47], and are notorious for their toxic behaviors such as racial slur, harassment, and personal attacks [50,72].

To deal with toxicity, Riot Games (Riot for short) has dedicated much effort into innovating behavioral systems. They actively promote sportsmanship through an honor system, pre-match messages, etc. [8]. They also maintained a participatory moderation system between 2011 and 2014 [54], as well as an automated moderation system thereafter [48,60]. The automated moderation system allows every player to "flag" others before or after a match, if they have exhibited toxicity of any form. Then the moderation system makes decisions about whether the reported behavior is toxic, and issues punishments accordingly. The punishments will escalate on an account based on the times of offenses [36]: 10 game chat restriction, 25 game chat restriction, two week suspension, and permanent suspension (or permanent ban).

Players who receive permanent bans usually receive an accompanying email containing their chat logs deemed toxic by the system. Riot does not provide official explanations about whether human labor is involved in the issuance of punishments. Nor has Riot released technical details about the system's automated decision-making process.

Riot does make several statements about permanent ban, or what we consider as Riot's discourse. Its TOS [62] writes:

> *7. USER RULES 7.1. Can I troll, flame, threaten or harass people while using the Riot Services? (No. If you do, we might take action such as banning your account.)*

Its support page [63] further explains:

> *The reality is that the behavior displayed was so negative or disruptive that it breached the Summoner's Code or Terms of Use. Such behavior can be severe enough to bypass all other warnings and result in the immediate banning of your account.*

Riot's discourse follows the basic logic of retributive justice, that a player receives PB as their punishment, and PB is deemed to be commensurate with the severity of the player's toxicity. Thus, the Riot discourse sees punishment as retribution (for toxic behavior).

Figure 1. Screenshot of a 5v5 match in League of Legends.

4 METHODS

In August 2020, the research team, composed of experienced LOL players, collected player discussions from the '/r/leagueoflegends' subreddit, one of the largest forums for LoL. Riot also uses the subreddit for occasional communication with the player community [53]. Our use of subreddit for discourse analysis aligns with previous HCI research [41,51]. Particularly, researchers have shown how online forums as a proxy for a game community could offer "a window onto underlying cultural logics and anxieties" [66].

We started by using the forum's search function to locate relevant discussion threads. Our initial keyword set was {permanent ban, permanent suspension, permanent account suspension}. But as we kept reading discussion threads and expanding the keyword set, our final keyword set included {permanent ban, permanent suspension, permanent account suspension, perma banned, perma'd, perm, perma ban, Riot ban, ban, punishment, punishment system, justice, penalty, and discipline.} For each discussion thread we encountered, we determined its relevance by completing an initial read to check whether there was at least a complete opinion or idea about PB. If PB was merely mentioned by players as they focused on other player experience issues, we deemed the discussion thread. Using this iterative search

strategy, we were able to locate 197 relevant threads with 7142 comments. The most commented thread had 539 comments, while the least commented thread had none. The size of the dataset is compatible with that commonly used in qualitative analysis [7].

Our data analysis approach combined thematic analysis [6] and discourse analysis [26]. And two coders were involved in this process. At the beginning, each of them assigned initial codes to all the ideas expressed in the dataset. The unit of analysis was usually a post or a comment where a player fully articulated their idea in terms of their arguments, reasoning processes, and evidence. At this stage, we used the big "D" discourse tool [26], to look beyond language and pay attention to the interlinks between language and people's beliefs, actions, interactions, values, and their social and institutional environment. For example, if a player expressed their criticism of permanent ban, we did not stop at assigning the piece of data as criticism. Instead, we looked for the justifications the player provided for their criticism, as well as the values players expressed through their language use. In doing so, we could locate the dispositions that players hold as they evaluate punitive systems. In addition, we could start to identify the discursive repertoire players commonly draw from the construct their speech acts. Here discursive repertoire consists of all the available discursive resources that players could use to build logics in their speech, such as their own experiences, community norms, practices, values, platform rules, and data published by Riot. This step allowed us to calculate the interobserver agreement (72.9%), which is considered good and nearly excellent [78]. After this step, we started to refine the initial list of ideas by comparing the similarities and differences between ideas. Through an iterative process, we combined similar ideas until we reached a satisfactory conceptual map, consisting of five player discourses.

We consider ethical implications of using available online data in this study. First, the university IRB approval was obtained prior to this study. In the IRB proposal, we discussed how our use of this online data would introduce little to minimal risk to LoL players, as the online data do not contain sensitive or personally identifiable information. Thus, we do not anticipate that any possible forms of harm to any LoL players. In addition, we believe the data could help provide unique insights into community moderation, which could be highly valuable for online platforms today. Second, given that the subreddit is open to the public, we sought to reduce the data's searchability by rewording them.

5 FINDINGS

Our analysis of permanent ban surfaced five distinct discourses. Next, we will describe each discourse, the discursive resources they draw from, and representative utterances.

5.1 Punishment as Retribution

This player discourse sees punishment as retribution for a toxic act. It aligns directly with the Riot discourse and ideas of retributive justice. One player wrote:

> *Riot is not a therapist. Their job is to enforce rules against the expectations they demand of their community... It is players' responsibility to handle their own frustration without abusing others. For example, if you come into my coffee shop and act aggressively against other customers, I'm allowed to blacklist you. The same is true for the game... You have to behave yourself or accept the consequences.*

This utterance draws parallels between the issuance of permanent ban (PB) and a real-world legal scenario, underscoring how the discourse borrows basic ideas from the Riot discourse and

thus the real-world legal system. According to this discourse, a player is expected to be self-contained and responsible for their deeds. PB is framed as a just response to an individual's toxic behavior, no more, no less. This discourse further legitimizes the losses PB incurs. Interestingly, the utterance does not just affirm Riot's authority in rule enforcement. By stating what Riot is not, it also intends to clear Riot of other responsibilities or possible improvements. Another player noted:

> *Permanent ban makes sense because it makes people lose something for their wrong doings. The players lose all the champions, as well as the money they have spent on the accounts. Temporary bans won't work because people don't worry about losing anything. Permanent ban means they have to do all the grinding work again.*

The utterance further explores the implications of PB on a banned player, in terms of losing access to game content including champions as well as cosmetic effects that players have already paid for. Such deprivation of privileges is believed to induce fear into players. Thus, PB also functions as a preventive mechanism in community moderation. The utterance acknowledges and justifies Riot's absolute jurisdiction over its game.

5.2 Punishment as Community Purification

The community purification discourse presumes punishment as retribution, like the first discourse. Additionally, it suggests a clear distinction between normal and toxic players. PB only targets the small set of malicious, irredeemable actors, and functions to purify the community. For instance, a player wrote:

> *Most people won't get permanent bans. Permanent bans are for players who ruin the game for the rest of the players. Those people are extremely toxic and want to abuse others at the very beginning of a match.*

The utterance above constructs a distinction between the "good" majority and the toxic few. Banned players are stigmatized as having problems inside their minds that are beyond fix. This is why banned players are either "extremely toxic," or have a toxic nature. Their toxic behaviors just take place naturally without external triggers. In a similar vein, another player wrote that "*permanent bans are so rare that one has to be a special jackass to get one.*" A third player wrote that "*Riot bans you because of your personality.*"

To further construct the difference between the normal and the permanently banned, this discourse also draws from a set of discursive resources that have been created by players or Riot. Here is an example:

> *Riot said those who were permanently banned were among the 0.06% most toxic players. They don't randomly ban people. Riot once gave banned players a chance to reform, but they reoffended again. Permanent bans exist for people who deserve it.*

The above utterance ascribes absolute authority to Riot over the adjudication of behavioral toxicity. From there, it cites two discursive resources to back up the claim. First, the percentage supplied by Riot creates the impression that banned players are neither many nor normal. They have a toxic nature, and they are the one bad apple that spoils the whole barrel. Second, the one-time reinstatement by Riot towards banned players further legitimizes Riot as a benevolent and reasonable authority that does have attempted to help banned players.

Because banned players are deemed to have a corrupt nature, they are beyond redemption and must be exiled from the community. A relevant quote is:

A permanent ban is not to reform you. A permanent ban means Riot doesn't want you in their game. They don't want your money or friends, because you are scum the whole time.

To reform is to assume that players are delinquents that could be corrected or reformed. But PB means Riot has determined the player's irredeemable nature. PB is justified as the only appropriate means to deal with such players.

But this community purification discourse is complicated by some banned players. One player wrote:

I have improved my behavior greatly in the past few years. I have been permanently banned twice. But I have been clean for several months now.

PB only applies to a player account, not the player who owns the account. Thus, the player in the above quote was able to create a new account to play the game. The above quote confirmed the community purification discourse in the sense that PB did eliminate a toxic account. But since the player came back, they were not really exiled from the community. But the player's quote claimed that their nature was purified: "Clean" is a word that many LoL players use to describe the reformed behavioral disposition of themselves, in a similar way used by a former criminal or a drug addict. As such, PB is viewed to have purified their minds by removing their contaminated parts.

5.3 Punishment as Platform Economy

The platform economy discourse stresses the game first and foremost as a commercial platform instead of a disciplinary place. As such, PB is positioned in terms of whether and how it benefits the profitability of LoL. The platform economy discourse is not focused on the legitimacy of PB, but the economic calculations behind it. Therefore, players who invoke this discourse could have different opinions towards PB. For example, a player reflected:

Riot loses a customer every time they ban someone. Therefore, this is the last measure they'd take.

The above utterance suggests a clear causal relationship, where a PB leads to the loss of a customer. Thus, PB is a poor business decision in terms of profits, and the last resort that Riot would turn to. While this utterances frames PB at the individual transaction level, other players consider PB's broader economic impact. Another player wrote:

Let's think of money-making and Riot's player base. Toxic behaviors could scare some potential customers off. This is not good for Riot. One bad player could ruin nine players' game at once, and cause them to quit the game. Permanent ban reduces toxicity. If the banned players create new accounts, they will likely spend money again in game, which still serves Riot's interest.

The player considered PB as a profitable measurement by Riot in terms of player engagement and retention. PB is framed as a way to protect the rest players' experiences and thus desirable. In this narrative, the underlining assumption is that player accounts, not players themselves, are associated with toxic behaviors. Therefore, the overall toxicity decreases if bad accounts are removed. Banned players can come back with a clean slate.

Yet not all players believe that a new account equates a reformed player. They are only sure about one thing, that is the economic benefit Riot can reap from a new account. A player wrote:

They permanently ban you if you are excessively toxic. But they will never ban your IP address or email address. Because if you cannot make a new account, they lose money. This is not good for the company.

The utterance acknowledges that it is a disciplinary decision to issue a PB, but it is a business decision to allow banned players to create new accounts, regardless of the players' toxicity. Thus, the utterance draws from two distinct set of discursive resources: Both discipline and profit.

The values of discipline and profit could be at odds and lead player conversations into a cynical direction. Here is an excerpt:

Riot should punish toxic players more. For example, they should be permanently banned after three strikes max. But Riot is a company and companies only want to make money. They just implement a chat filter, which is easy and cheap. They won't hire human reviewers because it's expensive. Riot's stance on toxic behaviors is nowhere near noble. They want you to think they do, but in fact they do nothing.

This utterance states that Riot's punitive system is ineffective in punishing toxic players, and inaccurate in detecting toxic behaviors. The suboptimal performance is motivated by economic reasons. Consequently, the current way of PB is insufficient and indecisive, reflecting Riot's inaction on the issue of toxic behavior. In a seemingly cynical conclusion, the player deconstructs the disciplinary system of Riot entirely, and criticizes it as being pretentious.

5.4 Punishment as Recirculation

The recirculation discourse disregards the disciplinary purpose of PB, and stresses PB's effect as the recirculation of toxicity. The assumption of this discourse is that banned players do not simply go away and are more likely to create new accounts and remain toxic. Consequently, toxicity is not eliminated, but simply reproduced and redistributed across the community. For instance, a player wrote:

Reporting abusive behaviors is somewhat meaningless despite what Riot says. If a player is banned, they could simply buy a new account for a small price and come back. Nothing stops the player from being toxic again. Money spent on an account is not really a major concern for many players.

The player cast doubt over the effect of Riot's moderation system, as well as Riot's discourse. As there is no public information about the number of banned players rejoining the game, the player was speculating about the general behavioral pattern for permanently banned players. In a way, uncertainty like the lack of necessary information is a discursive resource that players often utilize to make a point about the punitive system. Another player commented:

If permanent ban eliminates toxicity, why are we still experiencing it at a much greater scale? The truth is the justice system is broken... it issues unfair permanent bans, while ignores much worse toxicity.

Thus, the recirculation discourse tends to argue that PB has a cascading effect on toxicity: it redistributes toxic players. Here is a relevant quote:

If a player receives a permanent ban, it not always prevents the player from playing the game. I just had a game in Gold where we lost to a smurf from Diamond. He said he didn't want to smurf but his main account was banned. We had no chance in the game but had to endure...

This means that permanent bans punish other players like me... Are there other ways to allow toxic players to keep playing instead of ruining others' games through smurfing?

Smurf refers to the act of highly skilled player playing on newly created account, getting matched with new players or less skilled players, and ruining the latter's game experience. Smurfing is close to cheating and is considered toxic by many players. The above negative experience shared by the player demonstrates the valid side of the recirculation discourse, that when PB is applied to accounts, toxic players are not necessarily disciplined.

The recirculation discourse also criticizes the toxic disposition of people who smurf. A relevant quote is:

Permanent bans don't eliminate toxicity. Most banned players just buy another account and play again. They will be more toxic because it is not their own accounts. The system is not working.

The above utterance stresses the connection between toxic players and their accounts playing a role in the former's behavior, and hints at how such connection is ignored by the logic of PB. The above utterance suggests that affinity players could develop to accounts that they dedicate much time and effort to, which players do not have for new accounts. Therefore, banned players might consider newly bought accounts more disposable and show less care. Thus, new accounts are perceived to have a lower cost of committing toxicity.

5.5 Punishment as Black-Box

The black-box discourse criticizes the opacity of moderation and punishment. If the former four discourses focus on the goal or effect of PB, where the utterances suggest players know what PB does, this discourse reflects the unknowable quality of PB, where some players do not think they know what PB is. Players engage in this kind of discourse to question the decision-making processes behind the punitive system and call for more transparency in Riot's punitive system. Players trace the history of the moderation system to make a case for more transparency. Here is a conversation excerpt:

P1: In the past they just banned you without giving any explanation. Those were really dark times. Now they give us the chat logs, which is better... But I think it can still be improved. For example, they could highlight chat messages that the system thinks were offensive.
P2: Sounds great. But players might have more to complain to the support team. We know that Riot support never unbans anyone and will just encourage them to be more careful in game.

The two players in the above conversation have enumerated instances where punishments could be opaque, and sometimes unfair. Meanwhile, limited restorative effort is offered. The opacity could undermine the legitimacy of moderation. For instance, here is a conversation excerpt:

P1: I was banned, and I admit that I said bad things. But every game I came across players who were way more toxic.
P2: Riot is inconsistent with their bans. It's very frustrating that they don't allow much insight into their practices. People troll on purpose could get away, as long as they don't type.

The above conversation between P1 and P2 points to contradictions in players' moderation experiences. As those contradictions remain unresolved, players develop distrust in Riot's moderation system.

6 DISCUSSION

The discursive field of punishment contains diverse and sometimes competing player discourses. A player discourse encapsulates a particular group of dispositions and beliefs that players employ to interpret their personal experiences with PB. In the sustainment and construction of these discourses, players draw from a variety of discursive resources, such as Riot's official explanations and announcements, their personal experiences, and fellow players' experiences and perceptions.

Riot, the platform owner, is the most influential and only visible actor in the discursive field: two player discourses (a.k.a, retribution and community purification) overlap significantly with the Riot discourse, and all the five discourses concern the words or deeds by Riot. However, Riot's discourse does not dominate the field. The emergence of various player discourses is partly fueled by the disconnect between players and Riot, with instances such as the removal of player participation in its moderation system in 2014 [55] and the halt of its own forum in 2020 [25]. Player discourses echo such disconnect. For example, the community purification discourse consists of player speculations about Riot's disciplinary intention which does not appear anywhere in the Riot discourse. The black-box discourse questions the transparency of the punitive system and persists within the community, undermining the retribution discourse that claims PB as an effective means to combat extreme toxicity.

The vibrant and competing player discourses we found point to the diversity of player perceptions and experiences with punishments. The idea of paying attention to discursive field and discourse is not to assume a unified sentiment towards permanent ban in League of Legends, but to examine what is in the empirical data.

6.1 The Platform Rhetoric of Permanent Ban

PB represents the severest punishment for a player in LoL, and its ostensible objective is to discipline players into well-behaved community members who internalize and display sportsmanship. Discipline occurs along subjects' acceptance of certain knowledge that power seeks to impose into the them [22]. In this view, players do accept certain knowledge as truth: their discourses seem to firmly associate PB with the social category of "most toxic players." This knowledge matters because it carries a set of assumptions and statements, including establishing a distinct social category within the community, legitimizing PB as a form of punishment, and associating them together. The "most toxic player" would disrupt others' experiences on purpose and could be not reformed through any means. Thus, they deserve PB. In other words, PB does not work precisely to discipline players into well-behaved ones; It produces a stereotype of "most toxic player" within the community. The stereotype, in turn, justifies the platform's authority in issuing permanent bans.

But this does not mean it is successfully implemented and maintained as a disciplinary device. Punishment could be effectual through spectacle or surveillance [22]. Either way, public knowledge of a punishment matters greatly to its effect. But player discourses are also undermining PB's disciplinary effect to certain degree. First, there is not a clear behavioral threshold between PB and other less severe forms of punishment like chat restriction. The black-box discourse shows how players could not directly observe the logic behind the escalation of punishments but must base their understandings on indirect information such as what other players say on the subreddit. Second, PB seeks to exile a player in their entirety, assuming that the player is deemed beyond redemption. But the paradox lies in banned players' capacity to create or purchase new accounts and return to the community. Players also question

if game companies strive to maximize their profits by dataifying and commodifying punishments such as PB, much like the happenings on social media platforms [70]. They cast doubt on how Riot balances between managing toxicity and profiting through player engagement and retention, and, ultimately, the legitimacy and impartiality of Riot's moderation system.

High-profile PBs such as the Trump account suspension [14] are public rituals and have a rhetorical impact on the global discourse of free speech, legality, and corporate power which is beyond the conventional concerns of moderation. PB against an ordinary platform user, however, is largely invisible to public knowledge and remains a looming but mythic threat to individual users. It is not always predictable and almost never reversable.

Even when the platform claims to govern through punishments as severe as PB, players' various discourses challenge PB's status as the ultimate punishment, and thus render PB as platform rhetoric, a construct in the discursive repertoire that is simultaneously known and unknowable, powerful and ineffectual. Players' various discourses reflect PB's sometimes contradictory images within the community.

Discipline still happens, but for a different purpose. As the neoliberal economy disciplines people as individuals who must care for their selves [23], gamers are disciplined into individuals who freely make decisions about how to maximize pleasure from what games [32,33]. To understand the moderation system's disciplinary role, we must look beyond it and examine the orchestration of networked systems and practices that work together to structure player experience. It is true that the moderation system presents varied punishments as a threat. However, players already have created pathways not governed by the moderation system, such as creating/purchasing new accounts or devising toxic behaviors undetectable by moderation systems [44]. In addition, toxic meritocracy [61] disciplines players to accept values such as competition and achievement and to push aside other social and ethical values. In the toxic meritocracy of LoL, disciplined players engage with the ranking and matchmaking systems [47], keep playing and pursue higher ranks. Performance tracking systems turn players' attention to statistics that represent their game mechanics and knowledge [45]. In this environment, even if players are expected to perform sportsmanship, cooperating with teammates while respecting opponents, these expectations are largely secondary to the competitive ethos in LoL. Thus, some players would tolerate toxic but highly skilled teammates [46]. As such, players are still disciplined, but not into well-behaved community members. If the intended goal of moderation is to safeguard the community and manage toxicity, punishments in LoL have deviated from it.

We may expect a moderation system to engage players in pondering whether their actions adhere to certain behavioral standards. As rational agent of their own, the neoliberal player may question the legitimacy of the moderation system and only view it as a potential threat. To them, all punishments incur certain cost, and PB is associated with highest cost in terms of money and time to level up an account, nothing more, nothing less. Some players may view punishment as a calculated risk. They may develop knowledge about what kinds of toxicity the moderation system would or would not act upon [44], as well as routines to cope with PBs. As such, punishment may not simply reduce toxicity. It instead transforms toxicity into something crafty and skilled.

6.2 The Punitive Logic of Moderation and The Three Interlocking Elements of Toxicity (Behavior, Account, and Player)

Our findings and former discussion have pointed to the "wicked" side of the retributive justice rationale behind LoL's moderation system. The underlying assumption of the punitive logic produces a stereotype of toxic player, assuming a particular group of players, who are toxic players, create toxic accounts, and commit toxic behaviors. The assumption implies that they can reduce community toxicity as long as they could efficiently target and kick out those toxic players. The Riot discourse and retribution discourse reflect this conception. Such conception of toxicity has several ramifications. First, players feel that the moderation system fails to account for all types of toxic behaviors. In turn, they lose trust in the system. Second, players could not see the efficiency and performance of the moderation system but keep encountering toxic behaviors. In turn, they are more inclined to see the system as ineffective. Third, punished players are not satisfied with their experiences but cannot appeal. In turn, they create new accounts and remain toxic.

Reflecting upon the limits of retributive justice, HCI and CSCW researchers have turned to restorative justice [2,69] as an alternative lens to examine toxicity and moderation. Restorative justice refers to "the repair of justice through reaffirming a shared value-consensus in a bilateral process" [80]. For example, a mediation process could be arranged where the offender acknowledges wrongdoing, take responsibility, and express an apology. Most relevant to this study is the restorative lens could reframe toxicity and offender in LoL. Asad's comparison between the retributive and restorative lenses [2] is revealing in our findings: While the retributive lens defines offence in technical, legal terms, the restorative lens seeks to contextualize the offence, morally, socially, economically, and politically. In other words, the simplistic view of toxicity as intentionally malicious act is limiting. Existing literature suggests that toxic behavior is a prevalent phenomenon, could be contagious [12], and could occur on a "normal" player under immense frustration [44,77]. Our findings showed that all four discourses (except the retribution one) challenge this simplistic notion of toxicity to certain extent. The community purification discourse overlaps with the retribution discourse, but starts to see the line between toxic accounts and toxic players, on which the retribution discourse is not explicit. The platform economy one seeks to differentiate between toxic accounts and toxic players. The recirculation one extracts the idea of toxicity from concrete accounts or players. The black-box one directly questions the legitimacy of targeting toxic accounts.

From the restorative lens, to contextualize toxicity is to acknowledge three interlocking elements to approach toxicity: behavior, account, and player. A moderation system could be designed to eliminate toxic behavior, the account that exhibits toxicity, or the person behind the account. These goals are not mutually exclusive: A platform-wide message urging players to perform sportsmanship seeks to achieve the first goal without targeting any specific account. Punishments like PB applies penalty to a specific account, but players are free to create new accounts. Only in high-profile cases does Riot decide to ban the person behind the account permanently [63].

By unpacking toxicity along these three elements, we could analyze where the existing punitive logic breaks down. Riot claims that their punitive system targets only the most toxic players, presumably those who have high toxicity or frequent toxic behavior. However, in the actual execution, the punitive system only targets accounts that exhibit toxicity. Unsurprisingly, this leads to tensions between most of the player discourses and the Riot discourse. Table 1 summarizes these tensions.

Table 1. Tensions between Player Discourses and the Riot Discourse.

Discourses	Tensions
Retribution	N/A (Player discourse aligns with Riot discourse)
Community Purification	Whether PB eliminates toxic accounts or toxic players
Platform Economy	Whether Riot seeks to retain toxic players
Recirculation	What toxic players do next
Black-Box	How and why the system makes permanent ban decisions

Consequently, the retributive logic fails to capture the nuances of toxicity at different levels. First, by only targeting the accounts that have exhibited frequent toxic behaviors, the punitive system fails to account for the everyday toxic behavior by a "good" account that often erupts out of intense negative emotions [44]. Thus, the platform economy, recirculation, and black-box discourse would question the effectiveness of PB. Second, by only inflicting penalties on toxic accounts, the punitive system fails to account for the harm by toxic behaviors. Third, by only targeting toxic accounts, the punitive system fails to account for the players behind those accounts who are still toxic and keep returning to the community, as the recirculation discourse pointed out.

A restorative logic would consider all these three elements in a consistent framing. A productive approach is to consider ways to mitigate the tensions between the player and platform discourses. Given the dominant role of platform in the discursive field of moderation, the reconciliation between victim and offender is predicated first and foremost on the reconciliation between the user community and the platform. For example, the tension inherent in the community purification discourse rests on the simplistic assumption of toxicity, and could be resolved through community conversations. The recirculation and black-box discourse prevail due to the opacity of moderation design. These all point to the potentials of social and technical processes.

6.3 Implication for Design

Our findings point to the limits of the punitive logic behind LoL's current moderation system and indicate new possibilities for moderation design towards a more restorative approach. A restorative approach champions healing, reconciliation, and reparation [2]. To heal the wounds of victims, moderation design should consider several factors. First, moderation design needs to consider transparency as a mechanism to support toxic players. Previous research has demonstrated the helpfulness of transparency mechanism such as explanation associated with content removal in Reddit [39], our analysis of the black-box discourse also highlights how the lack of transparency could fuel criticism and distrust from players. Thus, it is important for platforms to put serious consideration to the transparency of their moderation systems. For example, they could provide more statistics about the ban rate, reform rate, and retention rate of banned players, as well as rationales behind moderation decisions. They should also be more transparent about whether and how players' reports impact the issuance of punishments. Admittedly, transparency efforts would demand more resources and investments from platforms. But the concern for cost should not be the only priority.

Second, moderation design should emphasize communication between victim and the platform. This is partially why the discursive field contains diverse and contradictory discourses. It appears that Riot is not seeking to communicate with its community. Its decision

to close its own forum (a.k.a, the boards) in March 2020 [25] also indicates the ignorance of the social and communicative aspect of the game. Thus, we call for more attention to building direct communication channels between platforms and their users.

Third, moderation design should rethink a restorative mechanism for victims. Schoenebeck et al.'s survey study [69] reviewed that victims of online harassment generally value traditional actions such as content removal and account suspension, but also attached importance to alternative approaches such as apology, payment, and offender list. Most importantly, they rejected the one-size-fits-all approach that may support some users at the expense of others. The key idea is to acknowledge the contextuality of toxicity and victim experience. Permanent bans are post-hoc acts that do not compensate for players whose games are already ruined, yet do not know if their reports have effects. Restorative mechanisms could be leveraged here to show care for those people who have experienced toxicity, in order to restore their trust in the platform. We already suggested in the above paragraph that transparency mechanisms allow people to feel their reports are effective. In addition, platforms could compensate for experiences of toxicity, if convicted. For example, video games could provide those people in-game gifts.

Our work points to the need to consider more participatory form of platform governance [18,68,71]. Our study showed how the player community and Riot are starkly different in terms of their perception of governance. As governance based on platform rules often fails to acknowledge the shifting community norms and dynamics [75], it would be helpful if platforms encourage participatory forms of governance, empowering user voices. For example, Wikipedia relies on editors for self-governance [20]. Previously LoL also encourages players to review and adjudicate on cases of reported players [49].

Lastly, our study also showed how punished players and particularly banned players are usually stigmatized and under-supported. They are on their own to navigate their permanent bans. Oftentimes they come to the subreddit, our study site, for suggestions from the community. If platforms want to retain banned players by allowing new accounts from them, then platforms should consider how to support behavioral reform on these players. For example, social venues could be designed where banned players could learn more about behavioral standards and what to do next.

6.4 Implication for Research

Our work points to several research directions in the future. First, while HCI and CSCW have seen much moderation research (e.g., [17,18,38]) in recent years, research on punishment experience is still scarce. More empirical research and conceptualization efforts should be carried out to reflect upon the punitive logic behind many moderation approaches and what are the restorative approaches to reconstitute the relationship between platforms and banned users.

Second, different from commonly studied platforms like Facebook and Twitter, our study site, an online game community, offers unique advantages in ruling out several confounding factors such as low barrier of entry and media attention. As such, our conclusions about punishment experience in LoL could yield translatable insights for studying that on other, social media platforms. For example, if a relatively closed community like LoL already has competing discourses about PB, how would Facebook users consider or even theorize about Facebook's punitive system? While certain discourses like retribution, platform economy, and black-box easily grow out of similar characteristics on social media platforms such as punitive system, surveillance capitalism, and moderation opacity, community purification and circulation

discourses may not when social media platforms employ real-person bans. All these point to more needs to investigate punishment experiences on other platforms, as well as cross-platform analysis for a fuller picture.

Third, our work is focused on permanent ban. We have briefly covered other forms of punishment such as warning and chat restriction. However, we did not discuss how these forms of punishment interrelate in player experience as well as how players experience the escalation of punishment. Future work could consider various forms of punishment as a whole. Survey study like [69] could be designed to surface users' preferences for punitive categories and alternative approaches such as apology and public offender list.

Fourth, our work is limited in analyzing only player discussion from a subreddit, which might skew the conclusions we have arrived at. Moving forward, other research methods such as interview and survey could be utilized to triangulate with online data.

7 CONCLUSION

In this paper, we examined discourses related to permanent ban in the League of Legends community. We analyzed permanent ban's rhetorical properties and disciplinary effect. Players' varied perceptions of punishment suggested that the existence of permanent ban was premised on a dated assumption that sees toxic behavior, account, and player as unified. This dated assumption leads to a false assumption that by eliminating toxic accounts, platforms could eliminate toxicity in their communities. This dated approach is further complicated by the market economy where platforms seek to retain their users. To manage toxicity, platforms should renew their understandings of toxicity and moderation.

ACKNOWLEDGMENTS

Many thanks to the reviewers for their cogent and constructive feedback, and Xinning Gui for her contribution to the qualitative coding process. The work is supported by NSF grant IIS-2006854.

REFERENCES

[1] Angela Adrian. 2010. Mischief and grief: virtual torts or real problems? *Int. J. Priv. Law* 3, 1/2 (2010), 70. DOI:https://doi.org/10.1504/IJPL.2010.029603

[2] Mariam Asad. 2019. Prefigurative design as a method for research justice. *Proc. ACM Human-Computer Interact.* 3, CSCW (November 2019), 41. DOI:https://doi.org/10.1145/3359302

[3] Hugo Adam Bedau and Erin Kelly. 2019. Punishment. *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*. Retrieved from https://plato.stanford.edu/archives/win2019/entries/punishment/

[4] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Human-Computer Interact.* 1, CSCW (2017), 24. DOI:https://doi.org/10.1145/3134659

[5] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (January 2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[7] Virginia Braun and Victoria Clarke. 2019. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*. DOI:https://doi.org/10.1080/2159676X.2019.1704846

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

[9] Edward Castronova. 2005. *Synthetic Worlds: The Business And Culture of Online Games*. University of Chicago Press, Chicago. Retrieved from http://www.amazon.com/Synthetic-Worlds-Business-Culture-Online/dp/0226096270

[10] Stevie Chancellor, Niloufar Salehi, Shion Guha, Sarita Schoenebeck, Jofish Kaye, Elizabeth Stowell, and Jen King. 2019. The relationships between data, power, and justice in CSCW research. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 102–105. DOI:https://doi.org/10.1145/3311957.3358609

[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Human-Computer Interact.* 1, CSCW (2017), 31. DOI:https://doi.org/10.1145/3134666
[12] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1217–1230. DOI:https://doi.org/10.1145/2998181.2998213
[13] Danielle Keats Citron and Benjamin Wittes. 2018. The Problem Isn't Just Backpage: Revising Section 230 Immunity. *Georgetown Law Technology Review*. Retrieved April 13, 2021 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3218521
[14] Kate Conger and Mike Isaac. 2020. Trump's Twitter Account Permanently Suspended. *New York Times*. Retrieved from https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html
[15] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media Soc.* 18, 3 (March 2016), 410–428. DOI:https://doi.org/10.1177/1461444814543163
[16] Kord. Davis and Doug. Patterson. 2012. *Ethics of big data*. O'Reilly.
[17] 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*. DOI:https://doi.org/10.1145/3290605.3300372
[18] 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*, 1–14. DOI:https://doi.org/10.1145/3313831.3376293
[19] Chek Yang Foo. 2008. *Grief Play Management*. VDM Verlag.
[20] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia Governance. *J. Manag. Inf. Syst.* 26, 1 (July 2009), 49–72. DOI:https://doi.org/10.2753/MIS0742-1222260103
[21] Michel Foucault. 1972. *The Archaeology of Knowledge*. Tavistock Publications Limited.
[22] Michel Foucault. 1977. *Discipline and Punish: The Birth of the Prison*. Vintage Books.
[23] Michel Foucault. 1988. *The History of Sexuality, Vol. 3: The Care of the Self* (First Vint ed.). Vintage.
[24] Christian Fuchs. 2012. The Political Economy of Privacy on Facebook. *Telev. New Media* 13, 2 (March 2012), 139–159. DOI:https://doi.org/10.1177/1527476411415699
[25] Riot Games. 2020. SAYING FAREWELL TO BOARDS. *LeagueofLegends.com*. Retrieved from https://na.leagueoflegends.com/en-us/news/community/saying-farewell-to-boards/
[26] James Paul Gee. 2014. *How to Do Discourse Analysis: A Toolkit*. Routledge. Retrieved from https://books.google.com/books/about/How_to_Do_Discourse_Analysis.html?id=O4SrAgAAQBAJ
[27] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media Soc.* 20, 12 (December 2018), 4492–4511. DOI:https://doi.org/10.1177/1461444818776611
[28] Tarleton Gillespie. 2018. *Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[29] Eric Goldman. 2021. Dear President Biden: You should save, not revoke, Section 230. *Bull. At. Sci.* 77, 1 (2021), 36–37. DOI:https://doi.org/10.1080/00963402.2020.1859863
[30] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. DOI:https://doi.org/10.1145/3173574.3174108
[31] James Grimmelmann. 2015. The Virtues of Moderation. *Yale J. Law Technol.* 17, (2015). Retrieved September 2, 2019 from https://heinonline.org/HOL/Page?handle=hein.journals/yjolt17&id=42&div=3&collection=journals
[32] Nicholas-Brie Guarriello. 2019. Never give up, never surrender: Game live streaming, neoliberal work, and personalized media economies. *New Media Soc.* 21, 8 (August 2019), 1750–1769. DOI:https://doi.org/10.1177/1461444819831653
[33] Renyi Hong. 2013. Game Modding, Prosumerism and Neoliberal Labor Practices. *International Journal of Communication 7*, 19. Retrieved September 15, 2020 from https://ijoc.org/index.php/ijoc/article/view/1659
[34] Sal Humphreys. 2008. Ruling the virtual world Governance in massively multiplayer online games. *Eur. J. Cult. Stud.* 11, 2 (May 2008), 149–171. DOI:https://doi.org/10.1177/1367549407088329
[35] Sal Humphreys and Melissa de Zwart. 2012. Griefing, Massacres, Discrimination, and Art: The Limits of Overlapping Rule Sets in Online Games. *UC Irvine Law Rev.* 2, 2 (2012).
[36] Itsumo. 2016. Instant Feedback System FAQ. *riotgames.com*. Retrieved from https://support-leagueoflegends.riotgames.com/hc/en-us/articles/207489286-Instant-Feedback-System-FAQ-
[37] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Human-Computer Interact.* 3, CSCW (November 2019), 1–33. DOI:https://doi.org/10.1145/3359294
[38] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput. Interact.* 26, 5 (July 2019), 1–35. DOI:https://doi.org/10.1145/3338243
[39] Shagun Jhaver, Amy Buckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Human-Computer Interact.* 3, CSCW (2019), 27.

[40] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput. Interact.* 25, 2 (March 2018), 1–33. DOI:https://doi.org/10.1145/3185593

[41] Charles Kiene, Donghee Yvette Wohn, Bryan Dosono, Kenny Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun "Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, and Kat Lo. 2019. Volunteer Work: Mapping the Future of Moderation Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing - CSCW '19*, 492–497. DOI:https://doi.org/10.1145/3311957.3359443

[42] Jason Koebler and Joseph Cox. 2018. Content Moderator Sues Facebook, Says Job Gave Her PTSD. *vice*. Retrieved from https://motherboard.vice.com/en_us/article/zm5mw5/facebook-content-moderation-lawsuit-ptsd

[43] Adam J. Kolber. 2009. The Subjective Experience of Punishment. *Columbia Law Rev.* 109, (2009). Retrieved March 15, 2020 from https://heinonline.org/HOL/Page?handle=hein.journals/clr109&id=186&div=7&collection=journals

[44] 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*, 81–92. DOI:https://doi.org/10.1145/3410404.3414243

[45] Yubo Kou and Xinning Gui. 2018. Entangled with Numbers: Quantified Self and Others in a Team-Based Online Game. *Proc. ACM Human-Computer Interact.* 2, CSCW (November 2018), 1–25. DOI:https://doi.org/10.1145/3274362

[46] 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*.

[47] Yubo Kou, Xinning Gui, and Yong Ming Kow. 2016. Ranking Practices and Distinction in League of Legends. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*, 4–9. DOI:https://doi.org/10.1145/2967934.2968078

[48] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Human-Computer Interact.* 1, CSCW (2017), 62. DOI:https://doi.org/10.1145/3134697

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

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

[51] Amanda Lazar, Norman Makoto Su, Jefrey Bardzell, and Shaowen Bardzell. 2019. Parting the Red Sea: Sociotechnical systems and lived experiences of menopause. In *Conference on Human Factors in Computing Systems - Proceedings*, 1–16. DOI:https://doi.org/10.1145/3290605.3300710

[52] Alex Leavitt. 2015. "This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 317–327. DOI:https://doi.org/10.1145/2675133.2675175

[53] Richard Lewis. 2015. A look at the relationship between Riot Games and the League of Legends subreddit. *dotesports*. Retrieved from https://dotesports.com/league-of-legends/news/riot-games-league-of-legends-subreddit-relationship-1606

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

[55] Lyte. 2014. Upgrading the Tribunal. *Player Behavior*. Retrieved from http://na.leagueoflegends.com/en/news/game-updates/player-behavior/upgrading-tribunal

[56] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 1138–1151. DOI:https://doi.org/10.1145/2858036.2858391

[57] Chip Morningstar and F. Randall Farmer. 1991. The Lessons of Lucasfilm's Habitat. In *Cyberspace: First Steps*, Michael Benedikt (ed.). MIT Press, Cambridge, MA, USA, 273–302.

[58] Casey Newton. 2019. The secret lives of Facebook moderators in America. *The Verge*. Retrieved September 3, 2019 from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[59] Arielle Pardes. 2020. All the Social Media Giants Are Becoming the Same. *Wired Magazine*. Retrieved from https://www.wired.com/story/social-media-giants-look-the-same-tiktok-twitter-instagram/

[60] Simon Parkin. 2015. A Video-Game Algorithm to Solve Online Abuse. *MIT Technology Review*`. Retrieved from https://www.technologyreview.com/s/541151/a-video-game-algorithm-to-solve-online-abuse/

[61] Christopher A Paul. 2018. *The Toxic Meritocracy of Video Games: Why Gaming Culture Is the Worst*. University of Minnesota Press.

[62] Riot Games. 2020. Riot Games® Terms of Service. *riotgames.com*. Retrieved from https://www.riotgames.com/en/terms-of-service

[63] RiotTyphon. 2019. Understanding Permanent Bans. *League of Legends Support*. Retrieved from https://support-leagueoflegends.riotgames.com/hc/en-us/articles/360038430973-Understanding-Permanent-Bans

[64] Sarah T. Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Stud. Publ.* (January 2016).

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

[66] Bonnie Ruberg, Amanda L.L. Cullen, and Kathryn Brewster. 2019. Nothing but a "titty streamer": legitimacy, labor, and the debate over women's breasts in video game live streaming. *Crit. Stud. Media Commun.* 36, 5 (October 2019),

466–481. DOI:https://doi.org/10.1080/15295036.2019.1658886
[67] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. 2018. Digital nudging: guiding online user choices through interface design. *Commun. ACM* 61, 7 (June 2018), 67–73. DOI:https://doi.org/10.1145/3213765
[68] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z. Tan, and Amy X. Zhang. 2020. Modular Politics: Toward a Governance Layer for Online Communities. *arXiv* (May 2020). DOI:https://doi.org/10.1145/3449090
[69] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media Soc.* (March 2020), 146144482091312. DOI:https://doi.org/10.1177/1461444820913122
[70] Ori Schwarz. 2019. Facebook Rules: Structures of Governance in Digital Capitalism and the Control of Generalized Social Capital. *Theory, Cult. Soc.* 36, 4 (July 2019), 117–141. DOI:https://doi.org/10.1177/0263276419826249
[71] Joseph Seering. 2020. Reconsidering Community Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Human-Computer Interact.* 4, CSCW2 (October 2020), 28. DOI:https://doi.org/10.1145/3415178
[72] Kenneth B. Shores, Yilin He, Kristina L. Swanenburg, Robert Kraut, and John Riedl. 2014. The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 1356–1365. DOI:https://doi.org/10.1145/2531602.2531724
[73] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *J. Child Psychol. Psychiatry.* 49, 4 (April 2008), 376–385. DOI:https://doi.org/10.1111/j.1469-7610.2007.01846.x
[74] Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung. 2017. Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. (October 2017). Retrieved November 2, 2019 from http://arxiv.org/abs/1710.04044
[75] Nicolas Suzor and Darryl Woodford. 2013. Evaluating Consent and Legitimacy Amongst Shifting Community Norms: an EVE Online Case Study. *J. Virtual Worlds Res.* 6, 3 (September 2013). DOI:https://doi.org/10.4101/jvwr.v6i3.6409
[76] T. L Taylor. 2006. Beyond management: Considering participatory design and governance in player culture. *First Monday* Special Issue 7 (2006).
[77] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. DOI:https://doi.org/10.1145/3313831.3376191
[78] Marley W. Watkins and Miriam Pacheco. 2000. Interobserver agreement in behavioral research: Importance and calculation. *J. Behav. Educ.* 10, 4 (2000), 205–212. DOI:https://doi.org/10.1023/A:1012295615144
[79] Chris Weedon. 1998. Feminism & the Principles of Poststructuralism. In *Cultural Theory and Popular Culture: A Reader*. Pearson Education, 172–184.
[80] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. 2008. Retributive and restorative justice. *Law and Human Behavior 32*, 375–389. DOI:https://doi.org/10.1007/s10979-007-9116-6
[81] 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*. DOI:https://doi.org/10.1145/3290605.3300390
[82] Bingjie Yu, Katta Spiel, and Leon Watts. *Supporting Care as a Layer of Concern: Nurturing Attitudes in Online Community Moderation*. Retrieved August 31, 2019 from https://www.metafilter.com
[83] Shoshana Zuboff. *The age of surveillance capitalism : the fight for a human future at the new frontier of power*.
[84] Mark Zuckerberg. 2018. A Blueprint for Content Governance and Enforcement. Retrieved from https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

Received January 2021; revised April 2021; accepted May 2021.