Democratic Moderation: Exploring the Use and Perception of Votekicking in Social Virtual Reality

Qijia Chen
Department of Computer Science
University of Helsinki
Helsinki, Finland
qijia.chen@helsinki.fi

Qunfang Wu
Harvard University
Cambridge, Massachusetts, USA
qunfangwu@fas.harvard.edu

Giulio Jacucci
Department of Computer Science
Helsinki Institute for Information Technology HIIT, University of Helsinki
Helsinki, Finland
giulio.jacucci@helsinki.fi

Abstract

Extensive research has focused on community-based moderation involving selected or volunteer moderators. The "votekick" system represents a democratized approach allowing all users to participate in moderation. Despite its widespread use in online gaming and social VR platforms, votekicking remains underexplored. This research studies how users use and perceive votekicking in VRChat, a leading social VR platform. Through thematic analysis of discussions from the Reddit community r/VRChat, our findings reveal that votekicking serves to cope with misconduct and enforce group-specific rules, but it also perpetuates toxicity such as materializing community-level biases. While praised for its immediacy and clear messaging against unacceptable behavior, votekicking's effectiveness is hindered by its reactive nature, consensus challenges, and decision-making complexities. This research contributes to broader discussions on the limitations and advantages of direct community involvement in moderation and suggests practical design improvements to address the challenges associated with votekicking.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; *Collaborative and social computing*; **Virtual reality**.

Keywords

social VR, virtual reality, community moderation, votekick

ACM Reference Format:
Qijia Chen, Qunfang Wu, and Giulio Jacucci. 2025. Democratic Moderation: Exploring the Use and Perception of Votekicking in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713577

1 Introduction

Online harassment is a widespread issue with profound consequences for the wellbeing of those targeted [115, 117, 126]. To combat such deleterious behavior, extensive research has investigated various content moderation mechanisms [17, 55, 56, 67, 107, 129]. Community-based moderation is one of the typical moderation strategies, which involves the active participation of users in the moderation process [42, 61, 106]. A significant body of research in Human-Computer Interaction (HCI) has explored community moderation practices across multiple platforms, such as Discord [68, 69], Reddit [18, 114], and Twitch [134]. In these environments, community moderators, often volunteers or selected members, enforce both sitewide guidelines and community-specific norms [63, 108].

The "votekick" system is commonly equipped in many online virtual communities. As the name suggests, combining "vote" and "kick," this system allows users to initiate and collectively vote for the removal or temporary ban of a disruptive user. This empowers the community to self-regulate and maintain a positive environment. The votekick mechanism democratizes moderation by allowing all members to participate directly in governance decisions. This makes a departure from typical community-based moderation which has been the focus of previous studies. Despite its prevalence, to the best of our knowledge, the votekick system as a moderation strategy remains largely unstudied across various online spaces. In particular, research investigating how votekicking is used and perceived as a governance tool is notably lacking, highlighting a significant gap in the literature.

We chose to study votekicking in the context of Social Virtual Reality (VR), where the "votekick" system is one of the most widespread governance mechanisms. Social VR is characterized by immersive 3D environments and lifelike avatars, enabling users to engage in real-time, embodied social interactions [87]. Social VR platforms transcend the limitations of traditional virtual spaces, offering enriched experiences and fostering deeper connections among users [80, 81, 122]. However, alongside its potential benefits, social VR also faces great challenges, particularly related to the prevalence of online harassment and toxicity [100, 110]. In addition, the uniqueness of social VR, including its immersive nature, strong sense of presence, and embodiment can intensify the psychological and emotional impact of harassment. Victims often experience harassment more acutely than in traditional social media, and new forms of abuse, new harassment forms (e.g., unwanted touch, spatial harassment) have emerged [41]. Social VR platforms adopt a wide range of anti-harassment governance mechanisms, including votekicking [142]. Given the prevalence and heightened harm associated with harassment in social VR, there is a growing call for research into the safety tools and moderation practices employed

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713577

on these platforms [24, 142]. Understanding how votekicking operates in such high-stakes scenarios is critical for refining these systems and ensuring user safety and well-being.

Researching votekicking can shed light on how empowering ordinary users to initiate and vote on moderation actions affects the overall moderation landscape. By examining user experiences and use associated with votekicking, researchers can assess its effectiveness, potential drawbacks, and general impact, providing a holistic understanding of this moderation strategy. In addition, understanding this is essential for more appropriate improvement, thus shaping a safer social VR environment and beyond. This research aims to fill the gap by examining the dynamics of the votekick function within VRChat [1], one of the leading social VR platforms. We pose two research questions:

- **RQ1)** What do users utilize votekicking for?
- **RQ2)** How do users perceive votekicking as a moderation strategy?

To address these questions, we collected user discussion data from the r/VRChat subreddit and then performed a reflexive thematic analysis of the data.

Our research contributes to the HCI community in the following ways:

- The study highlights votekicking's role in empowering community members to take immediate and democratic action against disruptive behaviors, distinguishing it from traditionally studied interventions. In addition, by showing how votekicking, as a democratic moderation strategy, serves as both a punitive and educational tool, the research highlights its role in informing and reinforcing community norms.
- Our research demonstrates the potential for votekicking to be misused, and underscores the inherent biases in user-driven moderation. This contributes to broader discussions on the limitations and fairness of direct community involvement in moderation, emphasizing the need for more informed and equitable strategies.
- The study identifies significant challenges in the voting process, such as user apathy and lack of informed deliberation, and compares these with other moderation methods, emphasizing the need for deeper user engagement and understanding.
- Finally, The study proposes practical design improvements, such as adjusting voting thresholds and enhancing user education, to mitigate votekicking's challenges.

2 Related works

This section provides an overview of the current state of research related to social VR and content moderation. We first examine the unique characteristics and challenges of social VR environments, particularly focusing on the prevalence of harassment and the moderation strategies employed to address it. Following this, we explore content moderation literature, distinguishing between platform-based and community-based moderation, and highlighting the role of votekicking as a democratic moderation tool.

[1] https://hello.vrchat.com/

2.1 Community-based Content Moderation

Online harassment significantly impacts the well-being of targeted individuals, posing threats to their mental and emotional health [115, 117, 126]. To combat such anti-social behavior, a large body of research has investigated various content moderation mechanisms to protect users from online harassment and ensure adherence to established standards, rules, or guidelines [17, 55, 56, 67, 107, 129]. These moderation strategies are crucial in maintaining a safe and respectful online environment. Content moderation on digital platforms can be broadly divided into two categories: platform-based moderation and community-based moderation [29]. In platform-based moderation, the platform itself oversees the enforcement of content standards through internal processes or automated systems [4, 7, 50, 58]. Conversely, community-based moderation involves users actively participating in the moderation process [42, 61, 106], aligning with principles of democratic decision-making. According to Saward, democratic decision-making contains representative democracy and direct democracy [90, 102]. Applying these forms to community-based moderation, representative democratic moderation is exemplified by community moderators, who are often volunteers or selected members tasked with decision-making responsibilities. Direct democratic moderation, on the other hand, is characterized by participation from all community members, achieved through voting, deliberative discussions, or other mechanisms that foster collective engagement. In the following, we review existing literature on the different forms of community-based moderation, guided by these democratic frameworks.

2.1.1 Moderation by Community Moderators. Content moderators are prevalent in online platforms such as Discord [68, 69], Reddit [18, 19, 114], and Twitch [15, 134]. Moderators enforce site-wide guidelines and community-specific norms [108]. These moderators are inherently part of the communities they serve, regularly participating in discussions and informal interactions [108]. As such, they must navigate a "dual identity" as both community members and authority figures. How moderators manage this balance can shape their conception of their role and influence their approach to moderation. Moderators contribute to the growth and development of the community by engaging new members, promoting community events, and fostering a sense of belonging among members. Their efforts in community development are critical for maintaining an active, vibrant community [91]. Research indicates that community moderators are often highly motivated by a desire to contribute to their community's growth and success. They strive to create engaging, dynamic spaces where members can connect, share, and learn from one another [63]. This sense of purpose is closely tied to the satisfaction of seeing their community thrive and become a valuable resource for its members [108].

Moreover, community moderators often acquire a deep understanding of the norms and values of their communities through ongoing engagement, which enhances their ability to moderate effectively [108]. This process of learning and adapting is described as "being and becoming" a moderator, where individuals gradually develop the skills necessary to perform moderation tasks and respond to challenges over time [29]. This development is seen as a form of reflective practice, where moderators continually assess and refine their strategies based on their experiences. By considering both the

immediate and long-term impacts of their decisions, moderators are able to adjust their approaches to better serve their communities [29]. This ongoing cycle of action, feedback, and reflection allows moderators to grasp the significance of their responsibilities and the profound impact they have on the health and dynamics of their communities [29, 135]. By maintaining the unique standards and implicit norms of their respective communities, moderators play a crucial role in fostering tailored, localized moderation efforts [20, 38].

The moderation practice, however, often involves considerable emotional labor from moderators, as they must navigate difficult content, mediate conflicts, and offer support to community members during challenging situations. This emotional toll can be demanding, requiring moderators to manage their own well-being while also assisting others [32, 116, 134]. In larger communities, volunteer moderators often face overwhelming workloads due to limited resources, leading to delayed moderation actions. This is particularly prevalent on platforms like Reddit and Facebook, where the scale of interactions frequently outpaces moderators' capacity to respond promptly [83]. Furthermore, community moderators often face criticism due to the perception that their decisions stem from power imbalances. When moderation authority is concentrated within a small moderator team, community members may interpret the process as corrupt, particularly when the decision-making is opaque or outcomes are unsatisfactory [136]. This perceived lack of transparency can lead to conflicts within communities [108]. Content editing and governance on platforms like Wikipedia have revealed oligarchic organizational structures [109]. As community moderators enforce site-wide governance rules and follow platform administrators' guidelines, online platforms often exhibit a form of "implicit feudalism," where governance processes reflect the values and interests of platform owners [103]. This centralization of power can lead to conflicts and a lack of trust, as the decision-making process typically involves little input from the broader user base.

2.1.2 Voting-based Moderation. In contrast to content moderators, direct democratic governance involved active participation from community members [75, 76], and in some cases, engagement in deliberation processes during moderation [118]. However, direct democratic moderation forms are less common than representative moderation by content moderators. A prominent example of direct democratic moderation is Reddit's karma-based system, which allows community users to vote on content. Users upvote or downvote posts and comments, and the resulting karma score reflects the number of upvotes minus the downvotes [98]. This system provides an accessible and lightweight mechanism for community members to express their preferences and influence governance outcomes[16]. Designed as a hands-off governance approach, the system encourages open dialogue and facilitates the sharing of controversial topics. Moreover, voting practices reinforce social identity within online communities, as users' voting behaviors often align with shared cultural norms and values [57]. Norm-following and prosocial factors are significant drivers for engaging in voting [95]. By amplifying contributions aligned with dominant community perspectives, voting mechanisms can elevate valuable and authentic information, especially during crises [74, 75].

Despite its advantages, voting faces notable limitations. Many users rely on others to vote, leading to widespread underprovision [46]. Additionally, concerns arise regarding the quality and fairness of voting due to biases and randomness in voting patterns. Positive voting exhibits a herding effect that inflates post scores, while negative voting shows inconsistent impacts [88, 132]. Most users do not engage deeply with the content they vote on, with many votes cast based solely on reading post headlines [47]. Users also reinterpret voting mechanisms based on their own norms and ethics, diverging from the platform's intended use [52]. More importantly, voting alone has been found to amplify harmful or exclusionary views and marginalize minority voices [44, 82, 137]. These shortcomings highlight the inadequacy of mere voting without deliberation. Research on democratic participation and deliberative democracy underscores the need for models that integrate informed dialogue and consensus-building to enhance governance legitimacy [33, 86, 93]. Fan and Zhang's work has demonstrated that deliberation provided a forum for collective reasoning and in-depth discussion; the deliberative model generally outperformed the blind-voting model but fell short in improving the efficacy of content moderation [34].

Beyond voting mechanisms, researchers have designed tools to facilitate direct democratic decision-making in content moderation. For instance, *PolicyKit* supports participatory governance by encoding community policies, enabling systematic and transparent decision-making [140]. To address issues of majority rules, Gordon et al. proposed a "jury learning" approach, which introduces controls over how majorities are formed and mitigate the risks of unchecked majoritarianism [49]. On Twitter, users can subscribe to curated blocklists that automatically block specific accounts [45]. Additionally, users can adopt customized feed algorithms developed by other community members to tailor their content streams [51]. While these tools often rely on third-party infrastructure, they mark a significant shift toward more decentralized and participatory moderation processes [66].

The votekick system is a form of community member-involved moderation commonly found in online gaming platforms and other virtual communities. While both votekick and systems like karma rating enable community-driven moderation, their mechanisms and objectives differ significantly. Karma-based systems, such as Reddit's upvote-downvote model, primarily aim to surface or suppress content based on collective community preferences. These systems often operate asynchronously and are designed for ongoing, granular evaluation of content relevance, quality, or alignment with community norms. The votekick system is action-oriented, focusing on the immediate removal or temporary suspension of disruptive users rather than evaluating content. This system democratizes governance by allowing all members to initiate and vote directly on user removal, making it a reactive tool for maintaining community norms and minimizing harm in real-time. Unlike karma systems, which aggregate preferences over time, votekick decisions often require rapid consensus and involve a higher degree of interpersonal accountability. Despite these distinctions, the votekick system remains underexplored in academic literature. To the best of our knowledge, the votekick system as a moderation strategy has not been systematically investigated. Existing studies only briefly mention votekicking or introduce the concept of the

function in online gaming or social VR context [28, 53, 78, 143]. For instance, the article When Bad People Happen to Good Games notes votekicking as a tool for addressing offensive behavior in a group setting [28]. However, no research has focused on understanding the use and user perceptions of votekicking in a comprehensive manner. By investigating the votekick system comprehensively, we aim to uncover its unique strengths and challenges, contributing to the broader discourse on community moderation.

2.2 Social VR and Harassment

In recent years, the domain of social VR has witnessed a significant surge in popularity. Social VR is defined as three-dimensional immersive digital environments that allow individuals to interact and communicate through the use of head-mounted displays [87]. Prominent platforms in this emerging field include VRChat and Horizon Worlds [2] [80]. In social VR, individuals are generally represented by avatars, which they control and animate using sophisticated body tracking technologies [40]. This technology supports lifelike interactions, offering verbal and non-verbal communication in real time [21, 81, 122]. In addition, the inherent technological features of social VR foster a strong sense of presence and body ownership [54, 127, 133].

With the above features, social VR is considered that transcend the capabilities of traditional virtual shared spaces [120], significantly enriching the social and experiential dynamics of online interaction. Previous studies on social VR illustrate people engaging in different social activities afforded by those platforms to meet a wide array of social needs [79, 120], from forging friendships and global networking to creative expression and digital gatherings for events like social dancing [94], sleeping [79], drinking [22] and mirror watching [23, 43, 123]. Such diversity in virtual socialization underscores the transformative potential of social VR as a medium for interpersonal connection and community building in the digital realm.

However, this new landscape is not devoid of challenges. Harassment remains a prevalent issue within social VR environments. Sabri et al. [100] conducted a virtual field study across three VR platforms, revealing that potentially harmful behaviors were observed in 45% of the events monitored. A survey conducted by Shriram and Schwartz [110] found that a significant portion of the users, including two out of seven women and 21 out of 99 men, reported experiencing harassment, with 42% witnessing harassment of others. Moreover, research indicates that the unique characteristics of VR technology, coupled with the diverse modes of communication encompassing verbal and non-verbal interactions, not only foster novel immersive experiences but also amplify the potential impact of online harassment beyond conventional text or voice-based forms [41, 81, 128]. The heightened sense of embodiment and presence in VR intensifies the experience of harassment compared to other computer-mediated social environments [41]. Additionally, the predominance of interactions between strangers in social VR applications contributes to increased conflicts [5].

In response to these challenges, recent research has delved into moderation strategies within social VR [25, 36, 105]. Schulenberg et al. [105] explored AI-based moderation, assessing its potential and limitations in creating safer online environments. This work underscores the need for nuanced and sophisticated AI moderators that can adapt to the complex dynamics of social VR. Another study by Fiani et al. [36] focused on the efficacy of embodied moderation agents, particularly in safeguarding younger users, highlighting the evolving landscape of moderation needs within these platforms. Furthermore, Zheng et al. [142] conducted an analysis to look at the types of safety features across social VR platforms, showing votekicking is one of the most adopted tools. This research advocates for a critical reassessment and redesign of those safety measures to better protect users and foster inclusive, respectful virtual communities. As the use of social VR grows, the need to evolve these tools becomes more apparent [3].

In summary, social VR offers immersive and natural social interactions that closely resemble those in the real world. However, these platforms are also plagued by toxic behaviors, which can have more intense psychological and emotional impacts compared to traditional social media environments. Tools like votekicking are commonly implemented to manage such issues, yet prior research highlights the need for further research to refine these mechanisms and improve user experiences. Despite its widespread use across social VR platforms, there is a notable gap in understanding how users perceive and utilize votekicking. Exploring how this tool operates in high-stakes situations is important for better ensuring user well-being and fostering a more secure VR environment. Studying votekicking in social VR not only addresses this gap but also provides insights that can inform the refinement of moderation practices in other real-time online contexts, such as online gaming and virtual meetings, which may be less complex yet share similar challenges.

2.3 VRChat as the Selected Platform

Social VR is accessible through various platforms, but our investigation specifically focuses on VRChat to study votekicking for several compelling reasons.

Firstly, VRChat is widely regarded as a leading social VR platform and serves as a flagship example of the potential of immersive virtual environments [92, 101, 138]. Its prominence ensures that insights from this study are highly relevant to a broad audience. Furthermore, VRChat consistently ranks among the top applications on the Steam store, highlighting its popularity and its ability to attract a diverse user base [80]. The heterogeneity of its users provides a broad spectrum of data, enabling a comprehensive analysis of user behavior and tool usage in virtual spaces, which is crucial for understanding votekicking practices. Additionally, VRChat features over 25,000 user-generated worlds [4] , fostering activities that range from casual hangouts and digital concerts to educational sessions and role-playing scenarios. This variety of social contexts provides a rich dataset for analyzing the use and perception of votekicking in diverse virtual interactions.

Another key reason for selecting VRChat is its well-documented challenges with harassment and toxicity. The platform has been described as a "wild west" due to frequent incidents of inappropriate

[2] https://horizon.meta.com/

[3] https://counterhate.com/blog/new-research-shows-metaverse-is-not-safe-for-kids/

[4] https://hello.vrchat.com/#:~:text=Over%2025%2C000%20Community%20Created%20Worlds%20and%20Growing

behavior [85]. For example, a study by the non-profit Center for Countering Digital Hate, which monitored VRChat activity for over 11 hours, reported numerous cases of behavior violating Meta's VR standards, including sexual harassment and abuse [1, 73]. The prevalence of such issues has drawn scholarly attention to VRChat as a space for studying harassment in immersive environments [92] and evaluating the effectiveness of safety tools [24]. This context makes VRChat an essential platform for exploring how votekicking systems address these challenges. Finally, the significant scholarly focus on VRChat, even outside the scope of toxicity research, reinforces its validity as a subject of study. Its importance within the field of social VR research is widely recognized, as evidenced by studies on diverse aspects of the platform [31, 43].

In VRChat, the votekick function allows players to remove disruptive or problematic users from a world through a voting process. This feature can be accessed in two main ways: Players can click directly on another user's avatar to open their profile. From the profile menu, there is an option to initiate a votekick against that player, as illustrated by Figure 2. Alternatively, players can open the nearby people list, which displays all the users present in the same world. From this list, a player can select any user's profile and start a votekick against them. Users do not need to provide an explanation for initiating a votekick. Once the votekick process is initiated, the system sends a notification to all players in the world, informing them that a vote is taking place to decide whether the targeted player should be removed. Each player is then given two options: "yes" (to vote for the removal of the targeted player) or "no" (to vote against it). For the votekick to succeed, a majority of players must vote in favor of the removal, referring to Figure 3. Once the voting period ends, all players who participated in the vote will receive a notification informing them of the outcome, whether the votekick passed or failed. If the votekick is successful, the targeted player will be removed from the world immediately. In addition to being removed, they also face a temporary block from re-entering that particular room for a set period of time, preventing them from disturbing the session again. Players can initiate multiple votekicks in a single world, targeting different users if necessary. For instance, if there are multiple users causing issues, separate votekicks can be started against each of them.

3 Research Method

In our approach, we collected discussions on votekicking from the r/VRchat subreddit, the largest VRChat community forum on Reddit. The collected data underwent a coding and analysis process conducted by two researchers.

3.1 Data Collection

Our method is chosen given our particular research questions, including understanding what users votekick and their perception of this tool. The inherent diversity within online forums offers a rich array of perspectives on votekicking, ranging from its applications to community perceptions, enabling a comprehensive analysis of mass discussions. In addition, online forums are often considered "safe spaces" for users to share their experiences and opinions freely and unsolicited [2]. Such environments allow for the collection of genuine insights. The r/VRChat subreddit, being one of the most significant social VR online community, serves as a prime place for natural exchanges among VRChat members. Reddit also acts as a valuable data source in CSCW and HCI research [22, 26, 71] for understanding users' behavior, perception, and experience, underscoring its suitability for our study.

To collect pertinent discussion data, we leveraged the official Reddit API, a tool facilitating keyword-based searches to extract content, titles of forum threads, and their corresponding comments. Focused on the specific keywords "vote" and "kick" due to their direct relevance to our study, we aimed to collect discussions pertinent to the votekick concept within the VRChat community. This initial extraction process yielded a dataset with 5,076 entries comprising 259 threads and 4,817 comments. After the initial data collection, we embarked on a data refinement phase, aimed at excising irrelevant and duplicated entries. While many posts mentioned keywords, not all were relevant to the votekick function, for example, the term "kick" can used in a body context, such as "I kicked him with my leg," Additionally, some threads touched upon multiple keywords, leading to their multiple inclusions in our dataset. To address these challenges, we first removed duplicate threads and associated comments, which left us with 233 threads and 4,381 comments. Then two researchers thoroughly reviewed the remaining threads to identify relevant discussions related to the votekick system. Irrelevant threads and their associated comments were deleted. Evaluating threads rather than individual comments streamlined the screening process, as the associated comments are likely to maintain a consistent theme with the threads. Cohen's kappa value among the coders is $\kappa = 0.82$, indicating nearly perfect agreement (0.81-1.00). Subsequently, then they held meetings to resolve any disagreements. The removal allows us to concentrate on discussions solely related to the social dynamics of votekicking in VRChat. After this refinement process, our dataset was distilled to 2,359 entries, including 103 threads and 2,256 comments. Collected posts span the time frame from 2018 to 2024.

Prior to initiating data collection, we thoroughly examined various ethical considerations, including adherence to the online community's norms, institutional research ethics, and broader ethical guidelines for online public data analysis. Initial steps included a careful assessment of the Reddit Data API Terms, ensuring our data gathering was in full compliance with community norms and did not infringe upon any subreddit guidelines. Our Institutional Review Board (IRB) deemed our study exempt from human subject review, recognizing our use of non-sensitive, publicly accessible data, in line with standard IRB protocols [97]. Despite this exemption, we remained cognizant of the ongoing ethical discourse within the research community regarding the utilization of public data [14, 37], with particular focus on issues related to privacy, anonymity, and the potential identification risks associated with online behavior. In response to these ethical concerns, we implemented several protective measures, including the removal of all personally identifiable information and the secure storage of collected data on a password-protected device accessible only to the research team. Furthermore, we meticulously rephrased all quoted material in our publication to diminish searchability and mitigate the risk of tracing content back to its original source. These proactive steps were taken to harmonize the valuable insights derived from analyzing publicly available

online discussions with our ethical obligation to safeguard the privacy and anonymity of the individuals whose data we examined.

3.2 Data Analysis

We employed reflexive thematic analysis [11–13] to analyze the data, which stands a further reflection of Braun and Clarke's widely used thematic analysis technique [10]. Reflexive thematic analysis fully embraces qualitative research values and recognizes researchers' experiences and pre-existing knowledge for developing, analyzing, and interpreting patterns critically. This analysis focused on discussion threads as the primary unit, ensuring the contextual integrity of each original submission and its responses was maintained. To achieve this, we amalgamated the title, the body of the initial post, and all related comments into a single document for each thread, identifying them uniquely via their post URLs. This methodological choice ensured that our analysis remained grounded in the specific discourse of the community while safeguarding the privacy of its members by abstaining from the use of individual-specific data.

Our analysis followed a structured series of steps [11, 27]. First, two coders independently reviewed the collected data to gain a thorough understanding of the content and context. This preliminary stage was pivotal for acquiring an in-depth grasp of the discourse surrounding votekicking within the VRChat community. After their independent reviews, the coders discussed their interpretations and resolved confusion regarding certain contexts in the dataset (e.g., What is "Persona 5"). Subsequently, the coders revisited the data individually for a detailed coding process. Text segments were coded based on their explicit meanings, with particular attention to comments elucidating the community's perception and application of votekicking. For instance, a comment stating "Votekicking is often used as a way to discriminate newcomers" was coded as "to discriminate newcomers". The next phase of our analysis focused on consolidating the initial codes into higher-level categories. This synthesis was key to distilling the data into meaningful themes that encapsulate the essence of the discussions surrounding votekicking in VRChat. The process was iterative, involving a continual back-and-forth between the dataset and the emerging themes to ensure they accurately represented the data. We then explored the internal relationships between the themes, merging categories with similar meanings to create a final set of overarching themes. Once this was done, we collaboratively refined and named the themes, carefully considering their essence to ensure they aligned with the data and research questions.

To help present our findings clearly and systematically, we have included two tables: Table 1 and Table 2 that display the identified themes and subthemes in an organized manner. We opted to forego inter-coder reliability metrics, a decision aligned with the qualitative research norms within HCI studies [84] and the principle of reflexive thematic analysis [11]. Our focus was on elucidating the categories of interest and exploring their interrelations to uncover broader patterns and insights into the social dynamics of votekicking within the VRChat community. This approach allowed us to delve into the complexities of community interactions, providing a nuanced understanding of the roles and perceptions of votekicking in virtual social spaces.

3.3 Positionality Statement

Positionality statements enable researchers to critically reflect on their own biases, identities, and assumptions [9]. By acknowledging these influences, researchers offer a more transparent view of their research process [59]. Within our research team, two out of the three authors have prior experience engaging with social VR users in previous studies and through personal interest. For instance, the first author has several years of experience interacting on various social VR platforms. All of the authors have conducted research related to online moderation, with one author having worked closely with community moderators in previous projects. The authors' experience makes us well-positioned to conduct reflexive thematic analysis in this study.

4 FINDINGS

4.1 What Users Votekick

The votekick function is a critical tool for enforcing conduct standards and maintaining thematic integrity, swiftly removing individuals who violate community norms or disrupt the collective experience. However, it is also used as a means for the purposes of harassment and discrimination.

4.1.1 Activities Violating Code of Conduct. The tool is employed as a measure against users who engage in harassment, indicating community-driven enforcement of conduct standards. Users who violate these norms, such as by using racial slurs or creating a disturbance, are subject to being voted out of the community spaces. For example,

> "I was hanging out with a bunch of people when suddenly an stranger walked in, repeatedly shouting offensive language and spamming sound effects... Shortly after, a request to remove him from the group pops up, and he was promptly removed."

This account demonstrates the use of the votekick function in addressing abusive incidents. By empowering users to take collective actions, the tool enables community governance.

The tool is also utilized to address other activities that violate community guidelines, such as user names that are unrecognizable, contain special or accented characters, or are otherwise against the VRChat community standards. For instance, *"So, earlier I witnessed some guy getting voted out. When I asked why, apparently he was not supposed to have a username like that, it wasn't allowed as it had bolded and darkened symbols."* The use of "bolded and darkened symbols" in the username is deemed inappropriate by the platform standards. [5]

[5]The use of special characters or symbols is prohibited by many online platforms for guarantees that usernames remain unique and easily recognizable across platforms; for example, Instagram enforces strict guidelines on the use of special characters such as @, #, and other in usernames. Here, the VRChat platform seems to follow a similar pattern to ensure easily recognizable and identifiable usernames. However, usernames with accented characters also play a role in user identification. These elements, alongside other features like voice, avatar, and user bio [104], can hint at attributes such as nationality or cultural identity. Unfortunately, this can lead to misuse, as users may exploit these identifiers to discriminate against others through mechanisms like votekicking.

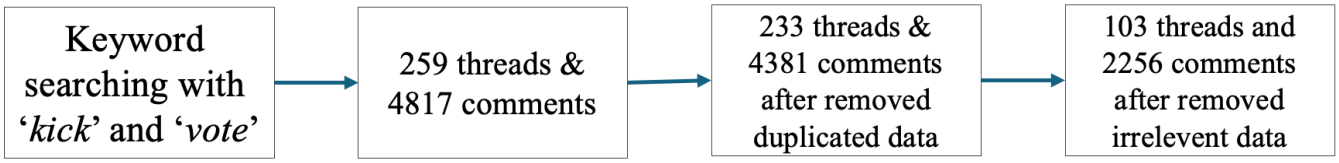

Figure 1: Data Collection and Refinement Process

Themes	Definition	Subthemes
1. Activities violating code of conduct	Users votekick those who engage in activities that violate guidelines, such as harassment, hate speech, gore avatars; Inappropriate user names.	Harassment; Hate speech; Oversexed avatars; Gore avatars; Inappropriate user names; Racist;
2. Actions Breaking Group Rules	Votekicking is used to enforce group rules such as maintaining thematic consistency.	Incoherent avatar; Inappropriate behaviors;
3. As a Means of Harming	The tool is exploited for harassment, disrupting users' experiences by removing them.	Kicking for entertaining; Kicking for threatening; Kick for disrupting; Kicking for no solid reasons;
4. Discrimination Against Specific Demographics	The votekick tool is used to discriminate against specific demographics, such as new users or those with particular identities or devices	For being newcomers; For being furries; Based on devices; Based on platforms; Based on race; Based on gender.

Table 1: Themes, definition, and subthemes for RQ1.

4.1.2 Actions Breaking Group Rules. Social VR platforms are often composed of diverse groups, worlds, and instances, each forming a unique microcosm within the broader virtual space. VRChat for example, consists of a diverse array of virtual worlds with different themes. Currently, there are over 25,000 worlds available [6] . Many worlds establish their own unique set of rules and moderation policies to ensure a cohesive experience for all participants. These policies are tailored to the specific needs and expectations of each community or group, covering a wide range of aspects from behavior to thematic presentation. The votekick tool serves as a mechanism for these communities to enforce their rules effectively, empowering users to maintain the intended atmosphere and integrity of their spaces. For example, in environments that are built around a specific theme—such as a map inspired by a video game, movie, or concept—avatar consistency can be a key rule. Participants are expected to choose avatars that fit the thematic nature of the space, contributing to a more unified experience. If not, the community may employ the votekick tool, thereby preserving the thematic integrity of the space. For example,

> "I joined the Persona 5 puzzle map, minding my own business. I wasn't bothering anyone or being disrespectful. As part of the map, you receive an anime character avatar. I switched back to my favorite avatar, and then suddenly, they started a vote to kick me out."

'Persona 5' is a 2016 role-playing video game developed by P-Studio and published by Atlus [7]. The Persona 5 world here had implicit and explicit rules regarding avatar use to enhance thematic immersion and gameplay coherence. By deviating from the avatar guidelines, even unintentionally, the user disrupted the immersive environment, leading the community to enforce the rule via a votekick. The response from another user further clarifies the underlying rationale for such community actions,

> "The essence of group instances is for each group to establish its own rules and moderation, within the guidelines of VRChat, of course. If you don't agree with the rules or atmosphere, it's best to find another instance or create your own."

In this context, the votekick function emerges not merely as a tool for conflict resolution but as a mechanism for subcommunities within VRChat to self-regulate and ensure their spaces remain true to their envisioned themes and rules.

4.1.3 As a Means of Harming. The votekick tool, designed to empower users to maintain community standards and ensure a respectful environment, can be twisted into a tool for harassing. The tool is sometimes exploited to harass users by threatening or actually kicking them out without valid reason, often as a form of entertainment or to disrupt others' engagement within the virtual environment. For example,

> "There was this one person blasting insanely loud music. So, I did the sensible thing and muted them, trying to carry on with my own business. But then they noticed I muted them and came charging over, making threatening gestures like running their finger across their neck as if they were going to decapitate me or something. After enduring this for a few minutes, I suddenly got sent back home with a message saying more than half of the users voted to kick me."

[6] https://hello.vrchat.com/#:~:text=Over%2025%2C000%20Community%20Created%20Worlds%20and%20Growing
[7] https://en.wikipedia.org/wiki/Persona_5

Blasting loud music is a common form of harassment in VRChat, often done using user-generated avatars equipped with sound elements. In this instance, the intended functionality of the votekick tool—to address genuine disruptions—is inverted. Instead of punishing the disruptive user, the tool was weaponized to penalize an individual who attempted to avoid conflict by muting the harasser. Another user echos the above statement from a more emotional perspective, further highlighting the misuse of the feature,

> "Yeah, it's frustrating when people abuse the vote kicking system just for kicks. I remember being in a similar situation once, chatting with a group of folks, and suddenly someone starts a vote kick for no apparent reason, and we couldn't even figure out who it was. It really ruins the experience."

The user shows frustration of being arbitrarily removed or threatened with removal. The misuse undermines the intended purpose of the tool and creates confusion and a sense of vulnerability among users.

4.1.4 Discrimination Against Specific Demographics. Another application of the votekick tool is in discriminatory actions against certain demographics. This includes targeting individuals based on their user status (e.g., newcomers), identity (e.g., furries [8]), device (e.g., Meta Quest users), or platform (e.g., Steam users). This moderation feature is highlighted as being co-opted to target individuals based on prejudiced grounds. Specifically, if there's an existing bias within the community against certain groups of people, the votekick mechanism can easily become a tool to further and materialize this discrimination. For example, new users face challenges in gaining acceptance within the community, who might be votekicked simply for being new, *"The kick could be due to you being new users (having a low rank). I've seen this happen a lot. It's strange, but some people seem to have an aversion to newcomers..."* New users are often unjustly targeted with votekicks. Previous research on social VR has shown community bias against newcomers [24]. Votekicking helps materialize the bias, hindering their integration and affecting their overall experience. Discrimination via votekicking not only ostracizes individuals but also contradicts the inclusive ethos VRChat strives for.

4.2 Users' Perception of the Tool

Votekicking is perceived as a potent moderation tool, offering immediate removal and a clear community message against unacceptable behavior. However, its effectiveness is challenged by, such as its reactive nature, difficulties in achieving consensus, and complexities in decision-making.

4.2.1 As a Powerful Tool. Votekicking is valued as a potent anti-harassment tool, offering immediate effects, sending a clear message to offenders, and imposing restrictions on their access to community spaces.

Immediate Effect and Clear Message. Users consider votekicking a powerful anti-harassment tool, as it sends a clear message to the kicked user about their unacceptable behaviors while having an immediate effect. For example, a user posts,

> "While I'm comfortable with using the blocking function, but it didn't seem like the right way to address a harassment situation. As it does not give the message to the toxic users that I want to give. Conversely, employing the vote kick feature against harassers can be an effective method..."

The user explicitly contrasts blocking, which "does not give the message to the toxic users," with votekicking, which they describe as "an effective method." This effectiveness arises from votekicking's ability to confront problematic behaviors both immediately and publicly. By addressing the offender as part of a collective response, votekicking holds the offender accountable in a way that private, unnoticeable actions like blocking cannot.

Imposing Consequence. In addition, the votekick tool not only removes users from the current space but also prevents them from rejoining for a certain period, introducing a consequential aspect of users' behaviors. The votekick mechanism is not just a measure for immediate action against disruptors or rule violators but also has relatively lasting implications. For example,

> "We had a case of a "Trol" votekick when someone went AFK, for an unusually long time. Upon their return, they were quite upset about being kicked... A votekick would prevent the person from rejoining for at least an hour."

AFK stands for Away From Keyboard in a social VR context, indicating that someone is not actively participating. By preventing re-entry for a set period, the mechanism demonstrates disruptive actions—whether intentional or perceived—have repercussions beyond the immediate moment.

4.2.2 With Limitations. Users perceived many limitations of the votekick mechanism within the virtual environments. These limitations span from the reactive nature of the system, difficulties in meeting votekicking conditions, to its susceptibility to exploitation, each contributing to the broader discussion on the effectiveness and potential for improvement in community self-moderation tools.

As a Reactive Tool. The reactive nature of votekicking means that it does not prevent harassment from occurring in the first place but rather addresses it after the fact. While it can remove a harasser from a game, it does not prevent the initial act of disruption. As a user shares, *"Tools like votekicking only work after the damage is done..."*

Conditions Hard to Meet. The conditions under which a votekick can be successfully executed are often not met. This issue is compounded in situations where a significant portion of the user base is unresponsive, making it challenging to gather the necessary support for a votekick. For example, a user posts, *"Why are vote to kick conditions never met. every time i try it, it does not work"* Another user echos the sentiment,

> "Some guy in our server is just screaming out slurs and overall being annoying but over half of the lobby is either afk or sitting in the corner of the world uunordered, it is difficult to kick them through the vote kick system."

These users complain about the difficulties posed by user apathy or inaction, which can render the votekick system ineffective even

[8] https://en.wikifur.com/wiki/VRChat

Themes	Subthemes	Definition	Subsubthemes
1. As a Powerful Tool	1.1 Immediate Effect and Clear Message	Provides immediate action and conveys community disapproval of unacceptable behavior.	Immediate effect; Clear Message; Reflecting collective disapproval
	1.2 Imposing Consequence.	Votekick removes users and prevents them from rejoining for a period.	Impose restriction; Having consequence; Stressing User Accountability
2. With Limitations	2.1 As a Reactive Tool	Votekicking only addresses harassment after it occurs.	Post-Incident Action; Not prevent harassment; Not prevent harm;
	2.2 Conditions Hard to Meet.	It is challenging to gather enough support to remove disruptive individuals.	Low success rate; Not enough people vote;
	2.3 Prone to Exploitation.	It is vulnerable to exploitation, leading to unfair removals and fostering a toxic atmosphere.	Easy to be used to harass; Easy to be used to discriminate; Easy to be misused; Lose faith in the tool
3 Problems with User Decision-Making	3.1 Impulsive Voting	Users tend to make impulsive decisions.	Convenience Over Conscientiousness; Lack of Engagement; Disregard for Consequences;
	3.2 Lack of Context	Users often lack the full context, leading to inappropriate voting.	Limited Grasp of Severity; Incomplete understanding of situation;
	3.3 Assumptions of Legitimacy.	Users often assume legitimacy of the initiatives of votekicking to default to vote	Default to Affirmative Voting; Default to vote no;
4. Backlash on Initiator	4.1 Initiator Criticism;	Initiators often face criticism, even having valid reasons.	N/A

Table 2: Themes, Subthemes, Definitions, and Subsubthemes for RQ2. User perceptions of the votekick tool demonstrate complexity, therefore we detail subsubthemes. Note that for Theme 4: Backlash on Initiator, no subsubthemes are identified and provided due to its straightforwardness.

in situations of clear misconduct. The tool's effectiveness is contingent on the collective engagement of users, which cannot always be guaranteed.

Prone to Exploitation. In addition, the votekick system's susceptibility to exploitation represents a significant flaw. Malicious users can manipulate the system to target individuals unfairly, leading to unjust removals from games or environments. This misuse not only undermines the intended purpose of votekicking as a tool for community self-regulation but also contributes to a toxic atmosphere within gaming communities. Such exploitation can discourage participation, foster mistrust among users, and ultimately degrade the overall user experience. Given we stated the point in section 5.1, we will not detail it again to avoid overlapping.

4.2.3 Problems with User's Decision-Making. This section can be considered a part of the limitations. However, we have observed that this part exhibits complexity and issues that we believe are very important to present separately, highlighting the nuanced challenges within the community moderation tool.

Impulsive Voting. Many users tend to prioritize convenience over conscientious participation. Upon receiving a voting notification, they do not take the time to understand the reason behind the vote and simply choose "yes" or "no", often to quickly dismiss the notification. For example,

> "When someone starts a kick, a notification pops up for everyone in the room. People see a button with "yes" or "no" and just click to get rid of the notification, without really thinking about what the vote is for. So, it looks like people are voting, but in reality, everyone is just trying to clear the notification."

The account shows when a votekick notification appears, the immediate goal for many becomes the swift removal of it, rather than an evaluation of the votekick's merits. A dialogue between two users further exemplifies this issue,

> "I don't care. I kick yes because it is easier [to make the popup disappear]" (A)
> "Do you not care about who's actually right though? What if a vote kick was against someone being bullied. You'd just say yes." (B)
> "Yes. Because I'm there to have a good time. If too much drama happens in an instance I have no problems bouncing to another instance. I view vrchat as a fun interaction. Nothing more." (A)

User A's stance represents a segment of the community that prioritizes personal convenience and immediate experience over the collective well-being and fairness of the community. The admission of voting "yes" indiscriminately, merely to expedite the removal of a popup, reflects a broader issue of responsibility in virtual spaces.

User B's questions attempt to probe the ethical considerations and potential consequences of such indifferent actions, highlighting the scenario where an unjust votekick could compound the distress of someone already being bullied. Despite this, User A's response underscores a preference for personal enjoyment and readiness to disengage from problematic situations by moving to another instance rather than confronting or resolving them.

The tendency to engage in superficial decision-making diminishes the effectiveness of votekicking and risks unjustly penalizing users based on the whims of those who vote without consideration. It highlights a disconnect between the intended purpose of the votekick system and the actual practice by some users within online communities.

Lack of Context. Users may not have the full context of the situation, leading to inappropriate decisions. This situation can be particularly problematic in virtual environments where interactions and conflicts may not be visible or understandable to all participants. Consequently, their votes might not reflect a fair judgment of the situation, potentially leading to the unfair treatment of individuals or, conversely, failing to address genuinely disruptive behavior effectively. A comment exemplifies the issue,

> "But honestly, most folks aren't really familiar with it. I've seen it happen a lot; they'll see the alert, but if, for example, you're busy chatting with people upstairs and the incident is downstairs, they won't vote appropriately because they don't have the full picture."

The virtual environments often feature dispersed activities and interactions. This user explains that users in one area may be unaware of conflicts occurring elsewhere. When votekick alerts pop up, users removed from the incident's location or not pay attention to the conflict are making decisions without a clear grasp of the context or the severity of the behavior in question.

Assumptions of Legitimacy. In many cases, users lack the context needed to make an informed decision and instead default to voting "yes" or "no" based on their trust—or distrust—in the system and the community. This behavior reflects a broader challenge in ensuring that the system operates as intended. For example, a user illustrates the default "yes" perspective,

> "Sometimes you can find yourself in such a huge world that you haven't even come across the group where the person being kicked is located. In those cases, I typically press the first option because there might be someone causing trouble and their group doesn't have enough members to kick them out."

This quote expresses the challenge of making informed decisions. The speaker acknowledges voting affirmatively ("the first option") without direct knowledge, assuming that the votekick is warranted. This reflects a belief in the collective ability of users to self-regulate and initiate votekicks only for valid reasons. Another user echoes this perspective, stating: *"Yes they can hit decline, but most assume kicking was initiated for legit reasons"*.

As highlighted previously, votekicking can be misused for harassing others rather than its intended purpose of community moderation. This has led some users to adopt a defensive stance, voting "no" by default, assuming the mechanism is more likely being exploited by harassers than used legitimately. For example, a user posts,

> "Most vote kicks are unwarranted. I do the same thing [decline the vote]. If I haven't seen what's going on, it's probably some troll trying to kick somebody for no reason. It's super annoying to get randomly kicked when you haven't done anything just because some troll decided to start a vote kick."

The user defaults to declining votekick requests, who does so out of a belief that the system is more often exploited than used appropriately. This defensive voting stance mirrors the frustrations of those who have been unfairly targeted by unwarranted votekicks.

4.2.4 Backlash on Initiator. The initiators of votekicks may face criticism or backlash, even when their reasons are valid. When someone takes the step to initiate a votekick, they are calling for a communal judgment on an individual's behavior. This act can be seen differently by community members: some view it as a necessary measure to maintain a positive environment, while others see it as an overreach or unnecessary escalation. The difference in perception can lead to discord and backlash against the initiator. For example,

> "There are instances where someone is either behaving extremely poorly or has been AFK in a gaming world for several hours, and yet, when I attempt to start a vote to kick them, it doesn't pass, even more frustratingly, people start bitch at me... I end up being the target for trying to initiate it."

The user shares their frustration that when they stepped forward to initiate a votekick, they are also putting themselves at the forefront of community judgment.

5 Discussion

This section delves into the dual nature of votekicking as a moderation tool. We first explore its empowering aspects, then examine its limitations. Finally, we propose design implications to enhance its efficacy and fairness. While this study focuses on votekicking in VRChat, its findings extend to other social VR and online gaming environments with similar moderation systems. Votekicking, as a moderation mechanism, generally involves user initiation and user voting (i.e., group decision-making). The benefits and challenges identified in the research —such as prompt responses due to the direct involvement of normal users, flaws in the voting process stemming from user apathy or limited contextual awareness, and issues with majority rule decisions—are not unique to VRChat but inherent in systems that rely on such moderation.

5.1 Empowering Communities Through Immediate Action and Collective Decision-Making

Our research advances the existing body of work on community moderation [15, 42, 68, 106, 134] by emphasizing the benefits of involving ordinary community members in initiating and making decisions about moderation. It highlights the superiority of votekicking as a moderation tool compared to other community

moderation strategies. In this section, we discuss two key aspects of votekicking: its ability to enable an immediate response to disruptive behavior and its role in educating users while reinforcing community norms.

5.1.1 Enhanced Immediate Response. Votekicking enables swift action against disruptive behavior, eliminating the delay often seen in moderation systems that rely on moderators or admins[108, 134]. Such delays can result in a backlog of unresolved cases and prolonged harm to the user experience [83]. Platforms like Reddit employ upvote/downvote systems as a real-time filtering mechanism based on community consensus. However, upvotes/downvotes primarily impact content visibility [52]. In contrast, votekicking removes the disruptive individual altogether, a more direct intervention tool. This is especially important in interactive environments like social VR, where real-time engagement defines the user experience. In addition, the immersive nature of social VR amplifies the impact of disruptive behavior [5, 6, 41]. Quick action is crucial to preserving the safety, enjoyment, and overall community atmosphere. Votekicking's ability to swiftly remove offenders helps maintain a safe and welcoming space for all participants. Moreover, votekicking's immediacy differs from other safety tools like blocking or muting, which only protect the user employing them and fail to address the harasser's impact on the wider community. In contrast, votekicking effectively eliminates disruptors from the shared space, preventing repeated harm and safeguarding the collective user experience.

5.1.2 Informing and Educating Present Users. Votekicking as a moderation tool extends beyond simple punitive measures, embodying a method for educating and influencing the broader community. It sends a clear message about the rules and expectations for everyone, including those watching and those directly involved.

Votekicking embodies a collective decision-making process by the community, rather than a top-down verdict from moderators. The process requires a consensus or a majority vote from the community, unlike moderation, where few individuals make the call [108, 134]. For users who are votekicked, this process potentially functions as a feedback loop, highlighting that their behavior has breached community standards. This form of peer-enforced moderation can be more impactful than an impersonal warning from an unseen moderator. Decisions made collectively are often perceived as more legitimate and fair because they reflect broader input [124]. Literature suggests that moderation outcomes from platforms can often be perceived as unfair or incomprehensible by the user [65, 70, 119]. Being removed by a visible voting process signals collective disapproval, potentially leading to a better personal reflection.

For users participating in the votekicking process, this act serves as an active engagement in the establishment and reinforcement of community norms. Collective decision-making fosters a sense of ownership and commitment to the decision outcome [64], as well as a better understanding of different perspectives of the situation (i.e., the right and wrong question of certain behaviors) [48]. Compared to decentralized systems like Reddit's upvote/downvote mechanism, votekicking amplifies the democratic nature of moderation by empowering users to take immediate action. Unlike systems using abstract indications like karma scores, votekicking excludes the users who misconduct, exerting direct influences on the community interactions and the experience. The intuitive outcome of votekicking potential serves as a powerful deterrent to potential disruptors, as it demonstrates that negative behavior carries immediate consequences. This visibility facilitates the communication and understanding of standards of behavior by community members. This contrasts with moderation where the outcomes might not be as transparent or widely shared. For instance, in competitive gaming moderation [70], punishments are often issued after the fact, when users are no longer present, leaving them uncertain about the consequences of misconducts.

5.2 Votekicking as a Moderation Tool that Often Fails

Our discussion in Section 5.1 highlights votekicking's potential as a robust moderation tool, yet it is crucial to acknowledge its significant limitations that challenge its efficacy and fairness. The discussion contributes to the broader understanding of community moderation [29, 63, 134], particularly on the pitfalls of direct, user-driven moderation. This section identifies and discusses three critical challenges associated with votekicking: issues in its initiation, flaws in the voting process, and problems arising from its reliance on majority rules.

5.2.1 Issues in Initiation. Votekicking is accessible to all users, which means they all can initiate votekick. However, many of them lack the knowledge or experience in content moderation, unlike those community moderators who have accrued expertise in managing community behavior and identifying misconducts [108] through their constant engagement and actively learning over time [29, 108].

This gap in knowledge, combined with the presence of malicious or toxic individuals, exacerbates votekicking's shortcomings. The universally accessible nature of the votekick tool, while enabling efficient and immediate responses to issues, it also makes it susceptible to abuse. Users can easily exploit votekicking to unfairly target individuals for personal vendettas, harassment, or amusement. In contrast, community moderators embody a different ethos, often bearing a strong sense of responsibility towards their communities [63]. They tend to have a strong desire to contribute to the development of their communities [63] and are motivated to see their community thrive [108].

5.2.2 Challenges in Voting Process. In addition to the issues of wrongly initiating the moderation tool, the votekick mechanism faces significant challenges in the voting process that undermine its efficacy.

Users show little interest in engaging with the democratic moderation process, viewing community maintenance as outside their scope of responsibility. This disinterest, compounded by the forced nature of participation in the systems, leads to apathetic voting behaviors. Instead of making thoughtful decisions, users may simply vote to get rid of the voting prompt. This contrasts sharply with the behavior of dedicated moderators, who are motivated by a sense of duty to actively manage and maintain their communities [29, 108, 135]. Votekicking's reliance on casual participation

introduces an inconsistency in decision-making, diminishing its reliability as a moderation tool.

The unique and complex nature of social VR environments adds further difficulty. These spaces are rich and multimodal, combining real-time verbal and non-verbal communication [81, 111, 122], which can make it difficult for users to fully grasp the context of an incident before voting. This starkly contrasts with platforms like forums or other social media, where evidence of harassment or toxic behavior is typically preserved in text form, allowing for clearer and more deliberate evaluations [68]. Moreover, the challenges in social VR exceed those found in voice-based harassment scenarios, such as in Discord [68], due to the presence of non-verbal harassment (e.g., unwanted touch and spatial harassment) [5, 81]. Such intricacies demand a nuanced approach to community moderation that acknowledges social VR's sophisticated dynamics.

The complexity also raises concerns about the legitimacy of votekick decisions. Voting alone without sufficient deliberation often falls short in establishing legitimacy. Participating in discussion activities improves the sense of procedural justice [141]. However, votekicking often lacks an informed deliberative process, which is essential for knowledgeable decision-making. This absence can lead to outcomes perceived as unfair, undermining trust in the tool's integrity. This highlights the importance of incorporating mechanisms that facilitate meaningful dialogue and collective reasoning.

The difficulty is compounded by voters' limited experience in identifying and interpreting harassment, as previously mentioned. This can lead to less knowledgeable voting, with many votes cast blindly, even if users have good deliberation. Research in deliberative democracy [96, 112, 121] identifies a key challenge: many participants lack the knowledge or skills for effective engagement. Our study extends this to democratic online moderation, where users' limited ability to identify harassment may lead to unintended harm. This lack of knowledge can lead to well-intentioned but ultimately harmful decisions, highlighting the need for education and guidance to enable fair and informed democratic moderation.

5.2.3 Limitations Due to Majority Rules. Beyond engagement issues, votekicking's reliance on majority rules based on users' votes also presents several problems, even though earlier discussions in 5.1 highlighted the benefits of democratic participation in determining punishments. One major problem arises when users exploit this system for personal or group gain. For instance, groups can conspire to protect their own members from being kicked or to unfairly target individuals who oppose them. This manipulation turns votekicking from a tool intended for community moderation into a weapon of exclusion, thereby intensifying community toxicity. This pattern mirrors a common flaw in democratic participation, where majority rule can marginalize minority voices, leading to what is referred to as the "tyranny of the majority." In such cases, the interests of smaller or dissenting groups are overshadowed by the dominance of the majority, further exacerbating imbalances in the community dynamics [62, 131]. Our findings on democratic online moderation reveal that organized user groups gain significant power under majority rule, often to the detriment of isolated or unaffiliated individuals.

Moreover, the dependence on majority consensus in votekicking can deepen existing biases within the community. A clear example is the bias against new users, who are often viewed with suspicion or hostility due to their unfamiliarity with established norms [72]. This predisposition makes it easier to vote newcomers out. In comparison, the karma system is not immune to similar issues of bias. Like votekicking, it can disproportionately penalize newcomers or unpopular opinions, often reflecting the preferences of dominant user groups rather than objective quality or behavior [44]. Organized brigades, much like groups exploiting majority rules in votekicking, can skew karma scores to promote their agendas or suppress dissenting voices [82]. However, the impact of such actions in the karma system may be less immediate and severe compared to votekicking, where decisions directly impact a user's ability to participate. This difference highlights a trade-off: while the karma system may fall short in addressing harassment effectively, its indirect nature reduces the potential for immediate harm, a limitation that ironically mitigates its overall impact. The challenges posed by majority rule in democratic online moderation necessitate a reevaluation of its application to ensure it serves its purpose of maintaining healthy community dynamics without fostering exclusion or reinforcing harmful biases.

5.3 Design Implications

Platforms that incorporate votekicking, whether in online games, virtual worlds, or other multiplayer environments, can benefit from the above findings by recognizing potential pitfalls and designing safeguards accordingly. Votekick systems are often misused, serving as tools for biases or unethical practices, or resulting in harm due to careless voting. Such misuse directly contradicts ethical design principles, which emphasize challenging biases that may manifest in designed systems and supporting the inherent worth and dignity of every individual [89]. Similarly, in the domain of online moderation, addressing bias and adopting strong ethical stances to protect vulnerable users are consistently emphasized [113]. These principles are equally relevant to votekick systems. To mitigate these challenges and enhance the efficacy of votekicking, the following design implications are proposed. We clarify that in addition to these immediate safeguards, moderation systems with ethical commitment require iterative improvement [3]. It is important to recognize the evolving nature of both technology and user behavior. Designers and developers need to continuously monitor the system's performance, collect feedback from users and stakeholders, and update the system to address emerging challenges and unforeseen consequences to ensure that the tool remains relevant.

5.3.1 Adjustment of Voting Thresholds. Adjusting voting thresholds can be effective. This strategy aims to ensure that the community can more effectively manage repeat offenders and prevent users with poor conduct from abusing the feature. Restricting rights is a common practice even in the physical world. For example, as a punishment for certain criminal activities – so-called criminal disenfranchisement [125] in most countries [99].

Taking into account a user's history of violations when determining the votekick threshold offers a layer of contextual sensitivity. Users who have repeatedly violated community standards might find themselves facing a lower threshold for future votekicks. This approach serves as a deterrent for repeat offenders. It signals that the community's tolerance for disruptive behavior diminishes

with each infraction. Conversely, users with a positive contribution history or no history of violations could benefit from a higher threshold, acknowledging their positive role within the community and offering protection against unjust votekick attempts.

The consideration of restricting voting rights to users based on their reputation or history within the community further refines the votekick process. Granting more weight to the votes of users who have demonstrated a consistent commitment to community values or giving them exclusive voting rights can help prevent abuse of the votekick system. This ensures that the power to influence community moderation lies with those who have a proven track record of positive contributions, thereby reducing the likelihood of votekicking being used as a tool for personal vendettas or harassment. Furthermore, to mitigate the negative effects of majority rule and safeguard minority voices, a potential design implication can involve assigning voting privileges to vulnerable populations and elevating their voting weights [49]. For instance, individuals belonging to marginalized groups based on their profiles or those frequently targeted by harassment can be assigned to voting tasks and given weighted votes.

Another refinement can involve limiting voting eligibility to users who are physically or virtually "close" to the individual being voted on. This approach prioritizes spatial and contextual relevance, enabling decisions to be made by those with direct knowledge of the situation [35]. For instance, eligibility could be confined to users interacting within the same physical location or virtual space. This can reduce uninformed or impulsive voting by distant users. Moreover, proximity-based restrictions may foster greater community trust, as members can be more likely to accept outcomes when they know the voters were actively present or involved during the incident in question.

5.3.2 Fostering More Positive Participation. Our research indicates that many users lack the motivation for votekick actions or vote carelessly. Encouraging a sense of belonging in decision-making increases engagement and responsibility. For example, schools often organize Parent-Teacher Associations (PTAs), where teachers and parents collaboratively discuss and establish school policies. This approach ensures that the policies reflect the needs of students' families and enhance parental involvement [139]. Similarly, online platforms can boost participation by involving users in creating or updating community guidelines through polls, suggestion forums, or collaborative drafting. For instance, polls should be conducted on the platform regarding what constitutes toxic behavior and what actions should be banned. This actively involves users and stimulates their public participation. Moreover, this is also an important means of promoting guidelines to users. Additionally, it is necessary to show users how their opinions have influenced decisions and those who contributed should be thanked. These measures can potentially increase users' involvement and hence more participation in votekicking.

In addition, user education plays a pivotal role in fostering responsible use. Regular reminders about the community's standards and the seriousness of votekicking can be integrated into the user interface at key moments, such as when initiating a votekick or reporting disruptive behavior. These reminders could deliver concise messages about the importance of upholding community values, the consequences of misuse, and the need to employ votekicking judiciously. By presenting these prompts at critical junctures, platforms can ensure users remain aware of the weight of their actions.

5.3.3 Providing Third Option for Non-Voting. Additionally, the votekick usually presents two options: yes or no. We believe it is very important to introduce a third option, "I don't know." to allow those who really do not want to vote or uninformed players from casting potentially harmful votes. This can draw parallels to real-life voting scenarios, which, although contextually different, still offer relevant insights. In most countries, voting during elections is not mandatory, with exceptions like Australia and Belgium [60]. One reason is that compulsory voting might force voters who are unprepared or lack knowledge about the candidates and issues to cast votes randomly, which may not accurately reflect the public's desires [30]. Moreover, voting is considered a right rather than a duty [77, 130]. In votekick, the default setting of only two options can be viewed as an implicit form of compulsory voting, which is why a third option should be available. Introducing a third "I don't know" option allows players to abstain without affecting the outcome. For example, if out of 8 voters, one chooses "I don't know," the decision whether to kick a player should be based on the remaining 7 votes. While this approach has advantages, such as protecting the integrity of the vote by allowing abstention, it also presents challenges. A potential downside is the risk of not achieving a sufficient number of decisive votes if too many opt for "I don't know." In these situations, we believe prioritizing fair and reasoned moderation is more crucial than simply obtaining a result. Accordingly, alternative strategies should be utilized to protect users.

5.3.4 Integrating Deliberative Processes. Votekick does not require users to provide explanations when initiating, leaving others with insufficient context to evaluate the situation. This contributes to the concern about the legitimacy of the voting outcomes. To address that, votekick could require users to provide a brief explanation when initiating a vote, assisting them in assessing the situation more effectively. Additionally, voting without sufficient deliberation often fails to establish procedural justice and can lead to poorly informed or blind decisions. However, direct democracy is generally more effective in smaller, cohesive groups [8]. As decision-making grows more complex, it becomes impractical for every participant to be fully informed on all aspects of the issue. Incorporating a lightweight deliberative process could enhance fairness without compromising efficiency. Fishkin [39] proposed a *Deliberative Poll* process—structured, time-bound discussions where key arguments are presented concisely. This deliberative approach enables users to deliberate effectively without excessive delay. This approach would balance the need for procedural fairness with the urgency required for real-time moderation.

5.3.5 Establishsing Safe Guides. To prevent misuse of votekick systems, it is also important to implement safeguard systems. For example, leveraging machine learning algorithms helps identify patterns of discriminatory behavior or misuse. The system could flag actions such as consistently targeting new users, specific demographic groups, or individuals based on device types.

These detection mechanisms serve as a line of defense to ensure the votekick feature is not weaponized against vulnerable groups or

used to perpetuate biases. Moreover, the flagged patterns can trigger further review by moderators or automated interventions, such as temporarily suspending votekick privileges for users or communities demonstrating problematic behavior. This proactive approach reinforces the system's integrity while upholding inclusivity for all participants.

5.4 Limitations and Future Works

The study's focus on discussions from the r/VRChat subreddit, while providing insightful observations, introduces limitations. The concentration on Reddit as a data source might not fully represent the wide array of experiences and viewpoints within the entire VRChat community, especially considering those users who tend to communicate through different platforms. Moreover, the unique community dynamics of VRChat may influence how the votekick tool is perceived, potentially limiting the generalization of the study's outcomes.

The study also opens avenues for future research to deepen our understanding and improve the application of these findings. Triangulating data through multiple sources, such as interviews with VRChat users or observational studies, could provide a more holistic and nuanced understanding of user experiences. In addition, future work could involve developing new technological solutions or modifications to the votekick system, drawing on research findings to create more equitable and effective moderation practices. Another interesting area is investigating how different levels of community engagement in VRChat influence the outcomes of vote kicking. For example, this could involve examining whether users with higher engagement levels experience more effective use of the vote kick tool due to better understanding and enforcement of norms to provide more understanding of community moderation. Additionally, future work could explore hybrid moderation models that incorporate proactive strategies, such as AI-assisted tools, to complement votekicking. AI could assist by detecting patterns of misconduct, initiating voterkick, and providing contextual analysis, or guiding players in making more informed voting decisions. Such a system could balance community-driven moderation with intelligent support, improving the tool's effectiveness.

6 Conclusion

Our research highlights the complexities of votekicking as a democratic moderation strategy, showcasing both its strengths and limitations Its immediacy is particularly beneficial in dynamic online spaces like social VR environments, where swift intervention is crucial to maintaining a positive user experience. Additionally, votekicking serves as an educational tool, helping to communicate and reinforce community standards. However, its universal accessibility also opens the door to misuse. Users can be unfairly targeted, and the voting process can suffer from user disengagement and the inherent complexities of real-time, multimodal environments. Majority rule can exacerbate biases, leading to the exclusion of certain individuals and reinforcing negative group dynamics. In response to these challenges, we propose design improvements to facilitate the better use of votekicking as a moderation tool.

Acknowledgments

We are grateful to the anonymous reviewers whose constructive comments have greatly enhanced the quality of our paper. This research received support from the Research Council of Finland (Grant #357270). Special thanks are also extended to Matti Nelimarkka for his valuable assistance and insights into the literature on democratic decision-making.

References

[1] [n. d.]. The Metaverse's Dark Side: Here Come Harassment and Assaults (Published 2021) — nytimes.com. https://www.nytimes.com/2021/12/30/technology/metaverse-harassment-assaults.html. [Accessed 18-11-2024].

[2] Nicholas Norman Adams. 2022. 'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19. *International Journal of Social Research Methodology* 0, 0 (2022), 1–16. doi:10.1080/13645579.2022.2111816 arXiv:https://doi.org/10.1080/13645579.2022.2111816

[3] Anders Albrechtslund. 2007. Ethics and technology design. *Ethics and information technology* 9 (2007), 63–72.

[4] Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. 2021. CyberPolice: Classification of Cyber Sexual Harassment. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*. Springer, 701–714.

[5] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[6] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social VR: Implications for design. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 854–855.

[7] Hannah Bloch-Wehba. 2020. Automation in moderation. *Cornell Int'l LJ* 53 (2020), 41.

[8] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. 2009. Voting in social networks. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 777–786.

[9] B. Bourke. 2014. Positionality: Reflecting on the Research Process. *The Qualitative Report* 19 (2014), 1–9.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[12] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[13] Virginia Braun and Victoria Clarke. 2022. Conceptual and design thinking for thematic analysis. *Qualitative psychology* 9, 1 (2022), 3.

[14] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4 (2002), 217–231.

[15] Jie Cai and Donghee Yvette Wohn. 2022. Coordination and Collaboration: How do Volunteer Moderators Work as a Team in Live Streaming Communities?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. Manipulating visibility of political and apolitical threads on reddit via score boosting. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 184–190.

[17] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[18] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 174 (nov 2019), 30 pages. doi:10.1145/3359276

[19] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[20] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and

macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[21] Qijia Chen, Andrea Bellucci, jie Cai, Matti Nelimarkka, and Giulio Jacucci. 2025. Understanding "Mutes" in Social Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3706598.3714244

[22] Qijia Chen, Andrea Bellucci, and Giulio Jacucci. 2024. "I'd rather drink in VRChat": Understanding Drinking in Social Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 797, 16 pages. doi:10.1145/3613904.3642405

[23] Qijia Chen, Andrea Bellucci, and Giulio Jacucci. 2025. Mirror Dwellers in Social VR: Investigating Reasons and Perception of Mirror Watching. *Proc. ACM Hum.-Comput. Interact.* 9, CSCW135 (April 2025), 23 pages. doi:10.1145/3711033

[24] Qijia Chen, Jie Cai, and Giulio Jacucci. 2024. "People are Way too Obsessed with Rank": Trust System in Social Virtual Reality. *Computer Supported Cooperative Work (CSCW)* (2024), 1–33.

[25] Qijia Chen, Seyed Mahed Mousavi, Giuseppe Riccardi, and Giulio Jacucci. 2025. Investigating the Use and Perception of Blocking Feature in Social Virtual Reality Spaces: A Study on Discussion Forums. *Proc. ACM Hum.-Comput. Interact.* 9, CSCW120 (April 2025), 19 pages. doi:10.1145/3711018

[26] Qijia Chen, Michiel Spape, and Giulio Jacucci. 2024. Understanding Phantom Tactile Sensation on Commercially Available Social Virtual Reality Platforms. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (2024).

[27] Victoria Clarke and Virginia Braun. 2021. Thematic analysis: a practical guide. (2021).

[28] Josh Coates. 2004. When Bad People Happen to Good Games: OK, so I admit it-not only am I a total closet gamer geek, I admit that I actually care enough to be bitter about it. Yep, that's right-this puts me in the "big-time nerd" category. *Queue* 1, 10 (2004), 92–91.

[29] Amanda L. L. Cullen and S. Kairam. 2022. Practicing Moderation: Community Moderation as Reflective Practice. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1 – 32. doi:10.1145/3512958

[30] Ruth Dassonneville, Fernando Feitosa, M. Hooghe, R. Lau, and D. Stiers. 2019. Compulsory Voting Rules, Reluctant Voters and Ideological Proximity Voting. *Political Behavior* 41 (2019), 209–230. doi:10.1007/S11109-018-9448-6

[31] Mairi Therese Deighan, Amid Ayobi, and Aisling Ann O'Kane. 2023. Social Virtual Reality as a Mental Health Tool: How People Use VRChat to Support Social Connectedness and Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

[33] J. Dryzek. 2001. Legitimacy and Economy in Deliberative Democracy. *Political Theory* 29 (2001), 651 – 669. doi:10.1177/0090591701029005003

[34] 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.

[35] Jeremy Ferwerda. 2015. *The politics of proximity: Local redistribution in developed democracies*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[36] Cristina Fiani, Robin Bretin, Mark McGill, and Mohamed Khamis. 2023. Big Buddy: Exploring Child Reactions and Parental Perceptions towards a Simulated Embodied Moderating System for Social Virtual Reality. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 1–13.

[37] Casey Fiesler. 2019. Ethical Considerations for Research Involving (Speculative) Public Data. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 249 (dec 2019), 13 pages. doi:10.1145/3370271

[38] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[39] James S Fishkin. 2018. Random assemblies for lawmaking? Prospects and limits. *Politics & Society* 46, 3 (2018), 359–379.

[40] Guo Freeman and Divine Maloney. 2021. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on human-computer interaction* 4, CSCW3 (2021), 1–27.

[41] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30.

[42] Dennis Friess, Marc Ziegele, and Dominique Heinbach. 2021. Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication* 38, 5 (2021), 624–646.

[43] Kexue Fu, Yixin Chen, Jiaxun Cao, Xin Tong, and RAY LC. 2023. " I Am a Mirror Dweller": Probing the Unique Strategies Users Take to Communicate in the Context of Mirrors in Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[44] Tiana Gaudette, Ryan Scrivens, Garth Davies, and Richard Frank. 2021. Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society* 23, 12 (2021), 3491–3508.

[45] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counter-public moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[46] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 803–808.

[47] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206.

[48] CK Gomathy, Mr V Datha Sai Bhaskar Sarma, and Mr V Manjunath Reddy. 2023. ORGANIZATIONAL BEHAVIOUR TRENDS AND DECISION-MAKING. (2023).

[49] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[50] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[51] Jay Graber. 2023. Composable moderation. https://bsky.social/about/blog/4-13-2023-moderation.

[52] Timothy Graham and Aleesha Rodriguez. 2021. The sociomateriality of rating and ranking devices on social media: A case study of Reddit's voting practices. *Social Media+ Society* 7, 3 (2021), 20563051211047667.

[53] Joanne E Gray, Marcus Carter, and Ben Egliston. 2024. Trust and Safety in Social VR: Current Industry Practices. *Governing Social Virtual Reality: Preparing for the Content, Conduct and Design Challenges of Immersive Social Media* (2024), 61–75.

[54] Scott W Greenwald, Zhangyuan Wang, Markus Funk, and Pattie Maes. 2017. Investigating social presence and communication with embodied avatars in room-scale virtual reality. In *Immersive Learning Research Network: Third International Conference, iLRN 2017, Coimbra, Portugal, June 26–29, 2017. Proceedings 3*. Springer, 75–90.

[55] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, 1 (2015), 68. https://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2

[56] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[57] Brent J Hale. 2017. "+ 1 for Imgur": a content analysis of SIDE theory and common voice effects on a hierarchical bidirectionally-voted commenting system. *Computers in Human Behavior* 77 (2017), 220–229.

[58] Qinglai He, Yili Kevin Hong, and TS Raghu. 2022. The Effects of Machine-powered Content Modereation: An Empirical Study on Reddit. In *55th Hawaii International Conference on System Sciences (HICSS)*.

[59] Andrew Gary Darwin Holmes. 2020. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education* (2020). doi:10.34293/education.v8i4.3232

[60] M. Hooghe and D. Stiers. 2017. Do reluctant voters vote less accurately? The effect of compulsory voting on party–voter congruence in Australia and Belgium. *Australian Journal of Political Science* 52 (2017), 75 – 94. doi:10.1080/10361146.2016.1260685

[61] 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 (2021), 1–24.

[62] Morton J Horwitz. 1966. Tocqueville and the Tyranny of the Majority. *The review of politics* 28, 3 (1966), 293–307.

[63] Jina Huh, R. Marmor, and Xiaoqian Jiang. 2016. Lessons Learned for Online Health Community Moderator Roles: A Mixed-Methods Study of Moderators Resigning From WebMD Communities. *Journal of Medical Internet Research* 18 (2016). doi:10.2196/jmir.6331

[64] Gred Islei and Geoff Lockett. 1991. Group decision making: suppositions and practice. *Socio-Economic Planning Sciences* 25, 1 (1991), 67–81.

[65] 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 (2019), 1–33.

[66] Shagun Jhaver, Seth Frey, and Amy X Zhang. 2023. Decentralizing platform power: A design space of multi-level governance in online social platforms. *Social Media+ Society* 9, 4 (2023), 20563051231207857.

[67] 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 (TOCHI)* 25, 2 (2018), 1–33.

[68] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[69] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[70] 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 (2021), 1–21.

[71] Yubo Kou, Yao Li, Xinning Gui, and Eli Suzuki-Gill. 2018. Playing with Streakiness in Online Games: How Players Perceive and React to Winning and Losing Streaks in League of Legends. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174152

[72] Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*. 11–20.

[73] Eli Cohen Lawson. [n. d.]. New research shows Metaverse is not safe for kids — Center for Countering Digital Hate | CCDH — counterhate.com. https://counterhate.com/blog/new-research-shows-metaverse-is-not-safe-for-kids/. [Accessed 22-11-2024].

[74] Alex Leavitt and John J Robinson. 2017. The role of information visibility in network gatekeeping: Information aggregation on Reddit during crisis events. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1246–1261.

[75] Alex Leavitt and John J Robinson. 2017. Upvote My News: The Practices of Peer Information Aggregation for Breaking News on reddit. com. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.

[76] Joon-Suk Lee, Seungwon Yang, Amanda L Munson, and Lusene Donzo. 2017. What people do on Yik Yak: analyzing anonymous microblogging user behaviors. In *Social Computing and Social Media. Applications and Analytics: 9th International Conference, SCSM 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II 9*. Springer, 416–428.

[77] A. Lever. 2010. Compulsory Voting: A Critical Perspective. *British Journal of Political Science* 40 (2010), 897 – 915. doi:10.1017/S0007123410000050

[78] Jonas Linderoth, Staffan Björk, and Camilla Olsson. 2012. Should I stay or should I go?–Boundary maintaining mechanisms in Left 4 Dead 2. In *Proceedings of Nordic DiGRA 2012 Conference*.

[79] Divine Maloney and Guo Freeman. 2020. Falling asleep together: What makes activities in social virtual reality meaningful to users. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 510–521.

[80] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Social virtual reality: ethical considerations and future directions for an emerging research space. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 271–277.

[81] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. " Talking without a Voice" Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[82] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society* 19, 3 (2017), 329–346.

[83] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[84] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[85] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping pro-social interaction in VR: an emerging design framework. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[86] David Moscrop and Mark E Warren. 2016. When is Deliberation Democratic? *Journal of Public Deliberation* 12 (2016), 4. doi:10.16997/JDD.256

[87] Fares Moustafa and Anthony Steed. 2018. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[88] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[89] Maurice Mulvenna, Jennifer Boger, and Raymond Bond. 2017. Ethical by design: A manifesto. In *Proceedings of the European Conference on Cognitive Ergonomics*. 51–54.

[90] Matti Nelimarkka. 2019. A review of research on participation in democratic decision-making presented at SIGCHI conferences. Toward an improved trading zone between political science and HCI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–29.

[91] Daniel Nielsen. 2022. Participatory Cultures of Digital Games: The Double-Edged Sword of Being a Reddit Community Moderator. *Baltic Screen Media Review* 10 (2022), 74 – 94. doi:10.2478/bsmr-2022-0007

[92] Laura Ortiz. 2022. Risks of the metaverse: A vrchat study case. *The Journal of Intelligence, Conflict, and Warfare* 5, 2 (2022), 53–128.

[93] J. Parkinson. 2003. Legitimacy Problems in Deliberative Democracy. *Political Studies* 51 (2003), 180 – 196. doi:10.1111/1467-9248.00419

[94] Roosa Piitulainen, Perttu Hämäläinen, and Elisa D Mekler. 2022. Vibing together: Dance experiences in social virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[95] Maria Priestley and Alex Mesoudi. 2015. Do online voting patterns reflect evolved features of human cognition? An exploratory empirical investigation. *PloS one* 10, 6 (2015), e0129703.

[96] M. Prisching. 2010. RATIONAL DEMOCRACY, DELIBERATION, AND REALITY. *Critical Review* 22 (2010), 185 – 225. doi:10.1080/08913811.2010.508633

[97] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society* 7, 2 (2021), 20563051211019004.

[98] Reddit. 2024. What is karma? https://support.reddithelp.com/hc/en-us/articles/204511829-What-is-karma.

[99] Brandon Rottinghaus and G. Baldwin. 2007. Voting behind bars: Explaining variation in international enfranchisement practices. *Electoral Studies* 26 (2007), 688–698. doi:10.1016/J.ELECTSTUD.2007.02.001

[100] Nazanin Sabri, Bella Chen, Annabelle Teoh, Steven P Dow, Kristen Vaccaro, and Mai Elsherief. 2023. Challenges of Moderating Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[101] David Saffo, Caglar Yildirim, Sara Di Bartolomeo, and Cody Dunne. 2020. Crowdsourcing virtual reality experiments using vrchat. In *Extended abstracts of the 2020 chi conference on human factors in computing systems*. 1–8.

[102] Michael Saward. 2003. *Understanding democratic politics: an introduction*. Sage. https://doi.org/10.4135/781446220962.

[103] Nathan Schneider. 2022. Admins, mods, and benevolent dictators for life: The implicit feudalism of online communities. *New Media & Society* 24, 9 (2022), 1965–1985.

[104] Kelsea Schulenberg, Guo Freeman, Lingyuan Li, and Bart Piet Knijnenburg. 2024. Does Who You Are or Appear to Be Matter?: Understanding Identity-Based Harassment in Social VR Through the Lens of (Mis) Perceived Identity Revelation. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–40.

[105] Kelsea Schulenberg, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, and Nathan J McNeese. 2023. Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[106] Joseph Seering. 2020. Reconsidering 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 (2020), 1–28.

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

[108] 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 (2019), 1417–1443.

[109] Aaron Shaw and Benjamin M Hill. 2014. Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication* 64, 2 (2014), 215–238.

[110] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*. IEEE, 225–226.

[111] Harrison Jesse Smith and Michael Neff. 2018. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[112] I. Somin. 2010. DELIBERATIVE DEMOCRACY AND POLITICAL IGNORANCE. *Critical Review* 22 (2010), 253 – 279. doi:10.1080/08913811.2010.508635

[113] Lucy A Sparrow, Ren Galwey, Dahlia Jovic, Taylor Hardwick, and Mahli-Ann Butt. 2024. Towards Ethical AI Moderation in Multiplayer Games. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY (2024), 1–30.

[114] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927.

[115] Simon Ståhl and Inga Dennhag. 2021. Online and offline sexual harassment associations of anxiety and depression in an adolescent sample. *Nordic journal of psychiatry* 75, 5 (2021), 330–335.

[116] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[117] Francesca Stevens, Jason RC Nurse, and Budi Arief. 2021. Cyber stalking, cyber harassment, and adult mental health: A systematic review. *Cyberpsychology, Behavior, and Social Networking* 24, 6 (2021), 367–376.

[118] Kil-Soo Suh, Seongwon Lee, Eung-Kyo Suh, Hojin Lee, and Jaehoon Lee. 2018. Online comment moderation policies for deliberative discussion–seed comments and identifiability. *Journal of the Association for Information Systems* 19, 3 (2018), 2.

[119] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13 (2019), 18.

[120] Philipp Sykownik, Linda Graf, Christoph Zils, and Maic Masuch. 2021. The most social platform ever? A survey about activities & motives of social VR users. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 546–554.

[121] R. Talisse. 2004. Does public ignorance defeat deliberative democracy? *Critical Review* 16 (2004), 455 – 463. doi:10.1080/08913810408443619

[122] Theresa Jean Tanenbaum, Nazely Hartoonian, and Jeffrey Bryan. 2020. " How do I make this thing smile?" An Inventory of Expressive Nonverbal Communication in Commercial Social Virtual Reality Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[123] Mattia Thibault and Mila Bujic. 2022. " Mirror Dwellers": Social VR, Identity and Internet Culture. In *DiGRA'22–Proceedings of the 2022 DiGRA International Conference: Bringing Worlds Together*. DiGRA online library.

[124] Tom R Tyler and E Allan Lind. 2001. Procedural justice. *Handbook of justice research in law* (2001), 65–92.

[125] Christopher Uggen, Angela Behrens, and Jeff Manza. 2005. Criminal disenfranchisement. *Annu. Rev. Law Soc. Sci.* 1, 1 (2005), 307–322.

[126] Jirassaya Uttarapong, Jie Cai, and Donghee Yvette Wohn. 2021. Harassment experiences of women and LGBTQ live streamers and how they handled negativity. In *ACM international conference on interactive media experiences*. 7–19.

[127] Vincent van Brakel, Miguel Barreda-Ángeles, and Tilo Hartmann. 2023. Feelings of presence and perceived social support in social virtual reality platforms. *Computers in Human Behavior* 139 (2023), 107523.

[128] Sara Ventura, Georgina Cardenas, Marta Miragall, Giuseppe Riva, and Rosa Baños. 2021. How does it feel to be a woman victim of sexual harassment? The effect of 360-video-based virtual reality on empathy and related variables. *Cyberpsychology, Behavior, and Social Networking* 24, 4 (2021), 258–266.

[129] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.

[130] Alexandru Volacu. 2020. Democracy and Compulsory Voting. *Political Research Quarterly* 73 (2020), 454 – 463. doi:10.1177/1065912919839155

[131] Benjamin Ward. 1961. Majority rule and allocation. *Journal of Conflict Resolution* 5, 4 (1961), 379–389.

[132] Tim Weninger, Thomas James Johnston, and Maria Glenski. 2015. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM conference on hypertext & social media*. 293–297.

[133] Carolin Wienrich, Kristina Schindler, Nina Döllinqer, Simon Kock, and Ole Traupe. 2018. Social presence and cooperation in large-scale multi-user virtual reality-the relevance of social interdependence for location-based environments. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 207–214.

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

[135] D. Y. Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). doi:10.1145/3290605.3300390

[136] Qunfang Wu and Bryan Semaan. 2023. " How Do You Quantify How Racist Something Is?": Color-Blind Moderation in Decentralized Governance. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–27.

[137] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. 2022. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–38.

[138] Shuiqing Yang, Kang Lin, Xi Wang, Yixiao Li, Yuangao Chen, and June Wei. 2024. Does personality matter? Understanding the impacts of real-self and avatar personality traits on metaverse satisfaction. *Internet Research* (2024).

[139] Denmark L. Yonson. 2016. Level of Parent Involvement in the Elementary and Secondary Levels. *The Normal Lights* 10 (2016).

[140] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.

[141] Weiyu Zhang. 2015. Perceived procedural fairness in deliberation: Predictors and effects. *Communication Research* 42, 3 (2015), 345–364.

[142] Qingxiao Zheng, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan C Salvi, Guo Freeman, and Yun Huang. 2023. Understanding Safety Risks and Safety Design in Social VR Environments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–37.

[143] Michael Zyda, Alex Mayberry, Jesse McCree, and Margaret Davis. 2005. How We Built a Hit Game-based Simulation. (2005).

A Pictures show the interfaces of votekicking in VRChat

Figure 2: This set of images, labeled A and B, shows two screenshots. They demonstrate the process of opening a player's user profile by clicking on their avatar and initiating a 'vote to kick' action.

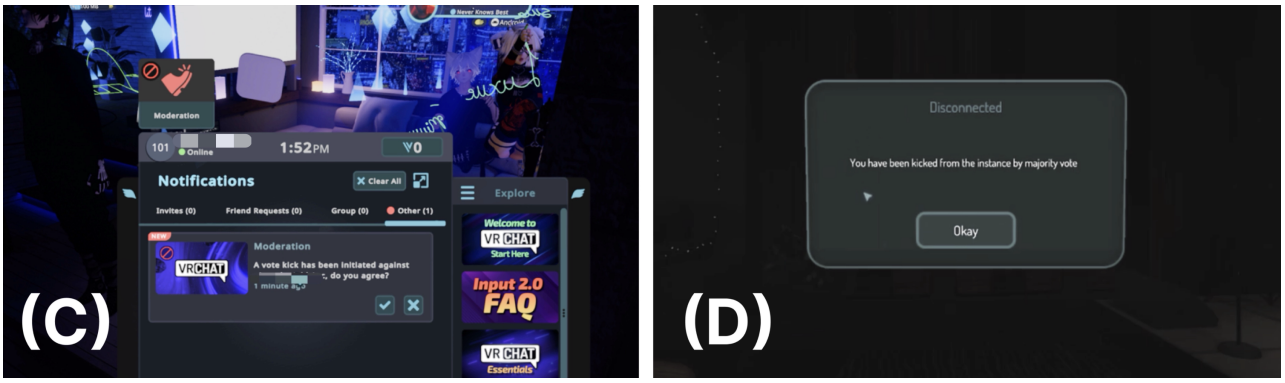

Figure 3: In image C, the player receives a system notification prompting them to vote 'Yes' or 'No' for a votekick. Image D displays the notification that the player has been kicked from the instance by majority vote.