UNIVERSITY OF CALIFORNIA,
IRVINE

Technologies of Social Safety and Moderation in Multiplayer Online Games - The Current State and Future Opportunities

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Ágnes Romhányi

Dissertation Committee:
Professor Constance Steinkuehler, Co-Chair
Professor Kurt Squire, Co-Chair
Professor Katie Salen-Tekinbaş,
Professor Daniel Epstein

2026

DEDICATION

To my daughter, Julia. You are my inspiration, my love.

To every young player online.

Have fun, stay safe!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My deepest gratitude to my family, my sister Judit, and my daughter Julia, for their relentless support, love, inspiration, and patience. This dissertation is yours as much as it is mine. I could not have accomplished this without you!

I am most grateful to my advisors, Constance Steinkuehler and Kurt Squire, for their guidance and encouragement throughout my graduate journey. It was a true privilege to be part of the Games+ Learning+ Society Center. Thank you for including me in many exciting projects and for supporting me in pursuing my interests. Working under your mentorship was invaluable; it broadened my views, honed my skills, and transformed my career.

I would also like to thank my committee members, Katie Salen-Tekinbaş, and Daniel Epstein, for the fruitful discussions and their assistance with my thesis and graduate journey.

Thank you to my fellow lab mates in the GLS Center: Minerva Wu, Jason Reitman, Maria Anderson-Coto, Richard Martinez, Lika Haizhou Liu, and most of all to Garrison Wells. I am infinitely grateful to you for all the support, for sharing the burdens, and for being my peers, mentors, occasional therapists, and most importantly, my friends.

Heartfelt gratitude to the brilliant undergraduate research interns for their work in collecting, cleaning, and organizing data, and for all the valuable insights and inspiring conversations.

The text in chapter three (3) is an adaptation of the paper as it appears in Agnes Romhanyi, Garrison Wells, and Constance Steinkuehler. 2026. Leave, Mute, or Retaliate? How Young Adults Handle Toxicity in Multiplayer Online Games. ACM Games Just Accepted (July 2026). https://doi.org/10.1145/3830080 used with permission from the Association for Computing Machinery (CC-BY 4.0). I kindly thank my co-authors, Garrison Wells for his assistance with piloting this study, data collection and analysis, and Constance Steinkuehler for her support and guidance throughout.

The Dissertation Writing Fellowship of the Informatics Department at the University of California, Irvine, provided financial support for this work.

VITA

Ágnes Romhányi

EDUCATION

Doctor of Philosophy in Informatics **2026**
University of California, Irvine *Irvine, California*

Master of Science in Informatics **2025**
University of California, Irvine *Irvine, California*

Bachelor of Science in Computer Science (Predegree Certificate) **2014**
Eotvos Lorand University *Budapest, Hungary*

Master of Art in Contemporary Museology **2014**
Eotvos Lorand University *Budapest, Hungary*

Master of Art in History **2008**
Eotvos Lorand University *Budapest, Hungary*

RESEARCH EXPERIENCE

Graduate Researcher **2020–2026**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2020–2026**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Romhanyi, A., Wells, G. and Steinkuehler, C. (2026) Leave, Mute, or Retaliate? How Young Adults Handle Toxicity in Multiplayer Online Games. https://doi.org/10.1145/3830080 **2026**

ACM Games Just Accepted (July 2026).

Wells G, Romhanyi, A., and Steinkuehler C (2025) Hate speech and hate-based harassment in online games. doi: 10.3389/fpsyg.2024.1422422 **2025**

Front. Psychol. 15:1422422.

Wells, G., Romhanyi, A., Reitman, J. G., Gardner, R., Squire, K., and Steinkuehler, C. (2024). Right-Wing Extremism in Mainstream Games: A Review of the Literature. https://doi.org/10.1177/15554120231167214 **2024**

Games and Culture.

REFEREED CONFERENCE PUBLICATIONS

Khan, F. R. and Romhanyi, A. (2023) Controlling Your Voice in a Shouting Match: A Preliminary Study on Fostering Self-Moderation among Gamers through Embodied Play. In Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play. Association for Computing Machinery, New York, NY, USA, 29–35. https://doi.org/10.1145/3573382.3616072 **2023**

CHI PLAY '23

Wells, G., Romhanyi, A., and Klaes, A,. (2024) Sus: Modifying Among Us for Misinformation Discernment. In Proceedings of the 19th International Conference on the Foundations of Digital Games. Association for Computing Machinery, New York, NY, USA, Article 58, 1–4. https://doi.org/10.1145/3649921.3656990 **2024**

FDG '24

ABSTRACT OF THE DISSERTATION

Technologies of Social Safety and Moderation in Multiplayer Online Games - The Current State and Future Opportunities

By

Ágnes Romhányi

Doctor of Philosophy in Informatics

University of California, Irvine, 2026

Professor Constance Steinkuehler, Co-Chair
Professor Kurt Squire, Co-Chair

Games are more than media for consumption; they are increasingly valuable educational tools and social platforms connecting players around the globe. The known benefits aside, harmful conduct prevails in online play, severely impacting players, gaming communities, and the games industry at large.

Players' understanding of social safety and their coping strategies with harmful conduct are influenced by the design and implementation of the safety infrastructure in online games. Therefore, improving these mechanisms can help reduce harmful behavior, enable proactive coping strategies, and develop healthier gaming environments. Safety infrastructure is a broader concept commonly incorporating policies, technological systems, and trust and safety teams. This work focuses on the technological components that directly shape player experience through structuring and moderating social interactions.

In three studies, (1) I assess parental controls and in-game tools of social safety and privacy, (2) examine how players perceive moderation systems, and (3) explore possible future avenues in AI-assisted moderation that move from common retributive approaches towards supportive interventions.

I argue that existing retributive moderation systems are insufficient to cultivate thriving communities, and that currently missing proactive and supportive measures are better aligned with player needs and shared expectations.

First, I show that retaliating against toxicity or leaving it unchallenged solidifies toxic norms; yet to a lesser extent, victimization and game experience also foster resilience and productive coping. Proactive tools that help players respond with constructive strategies rather than reciprocating harm are, therefore, pertinent.

I use the Teens Online Safety Strategies (TOSS) framework to review parental tools and social safety controls in popular multiplayer online games and publishing platforms. I demonstrate that teens' developmental needs are not well supported, and parental systems both infantilize them and leave them vulnerable. The available social safety and privacy controls are mostly restrictive, and mechanisms supporting active coping and impulse control are scarce, if they exist at all.

The recent surge in utilizing AI invites the question of its potential for innovating moderation systems. Thus, I explore players' attitudes toward real-time AI interventions aiming to reduce toxicity and foster prosocial behavior. I elucidate that the most opportune solutions seamlessly integrate AI-assisted mediation into gameplay and follow strong ethical considerations for transparency, privacy, as well as human agency and control.

This dissertation delineates the current state of safety infrastructures and technological moderation in multiplayer online games and specifies viable future directions that may inspire a cascade of work exploring proactive supporting ways of moderation.

Chapter 1

Introduction

1.1 The Context

Video games have become a leading medium for digital entertainment, socializing, and learning. Today, the video game industry generates worldwide revenue that exceeds that of both the film and music industries [100]. In 2024, according to the Entertainment Software Association (ESA), more than 190 million Americans played video games regularly [60]. Over the past two decades, multiplayer online games (MOGs) have experienced rapid growth in popularity. The percentage of those who reported playing online in the US increased from 18% to 88% between 1999 and 2023 [60].

Video games have numerous benefits. They have been found to improve cognitive skills, such as visual reflexes, multitasking [181], problem-solving, and spatial skills [50, 209]. Multiplayer online games encourage prosocial behavior [208] and collaboration [77, 86, 87], effectively boost players' online social capital and civic engagement [265] and create fun and rich learning opportunities [174, 220]. Furthermore, online games serve as third places for adolescents to socialize and encounter diverse ideas and worldviews beyond their homes, schools, or

work [223].

However beneficial, participation in networked games is not without risks. With the growing ecosystem of online games, there are increasing technological and social threats to player safety. Although technological risks, such as viruses, malicious software, insecure game servers, and vulnerable game code, are consequential, they are outside the scope of this dissertation. This work focuses on the social harms that arise from player-to-player interaction and explores infrastructures of social safety, privacy, and moderation systems that govern player behavior. [1]

Harmful conduct, or more commonly toxicity, is a well-known dark side of online gaming. It is an umbrella term for various verbal and behavioral transgressions [2, 207, 235] that adversely impact player experience [82, 125, 130, 137], and cause significant mental, emotional, or even physical harm to the players and others in their lives [62].

Common examples of toxic behavior include incivility, degrading comments [206], mischief, cheating [2, 114], griefing, trolling, and harassment [2, 233, 266]. More egregious but perhaps less prevalent forms of toxicity are cyberbullying [9, 26, 42, 75, 107, 176, 189], hate speech, and hate-based harassment [240, 243, 269], child sexual abuse, and predatory conduct [53, 63], as well as extremist propaganda [138, 200, 242].

In their report *Hate is No Game: Hate and Harassment in Online Games 2023*, the Anti-Defamation League (ADL) found that in 2023, 76% of adults and 37% of children experienced harassment in online games [7]. Similarly, the Pew Research Center found that in 2023, 43% of teens (ages 11 to 17) have been harassed or bullied while playing online [98]. Members of marginalized groups, who do not fit within the dominant heterosexual masculine gamer identity, are the most frequent targets of harassment [9, 11, 68, 182, 213], along with novice

[1] System, mechanism, tool, and intervention are used interchangeably, and when discussing moderation, I refer to technological apparatus. Although human moderators, community managers, and other staff are crucial in the safety infrastructure, this dissertation focuses on the technologies.

players who often face skill disparagement [206, 266].

Harmful conduct has multiple negative ramifications for individuals, social play, and the gaming community at large. It causes embarrassment, anger, fear, lowered self-esteem [48, 201], diminished self-efficacy, loneliness [75, 201], and in more severe cases can lead to drug abuse, depression [85], and suicidal thoughts [92].

In team-based multiplayer online games, toxic behavior is more commonly rooted in performance anxiety and emotional dysregulation [255] and often directed at teammates rather than opponents [206, 256].

Exposure to toxic behavior reinforces its acceptance, which in turn leads to engaging in harmful conduct more easily [11, 91]. Those who play more and for a longer time are more likely to become perpetrators [11, 206], and victims often respond with matching hostility, creating a vicious cycle of harm [70, 235]. Thus, replicating toxic behaviors, gaming communities play an active role in normalizing toxic practices in online games [11, 206].

1.2 The Research Problem

Although toxicity in online games is part of larger social issues, player behavior, and thus harmful conduct, are shaped by the technological systems that define the boundaries of social play. These technologies include the rules embedded in game mechanics and meta-systems, communication affordances, safety and privacy features, and moderation tools. The design of these systems has a profound impact on how players interact with one another, the behaviors and play styles they develop and recognize as successful, and how they understand and manage online safety or respond to toxicity.

To protect their players, game companies developed a complex technological infrastructure, including age restrictions, parental controls, safety and privacy features, moderation tools,

and other proactive interventions aiming to foster prosocial behavior. Yet, they struggle to build thriving online communities. These systems are at the center of my investigation.

Prior work highlighted several shortcomings of the safety and moderation infrastructure. Many of these tools fail to engage players and remain underutilized; about half of the parents don't set up parental controls on their children's accounts [67], and players are largely unaware of privacy and safety settings in games [15]. Other systems, such as reporting, are often misused and considered ineffective by the players [11, 36, 127, 129, 257].

Moderation systems are also commonly criticized for excluding players and following a top-down retributive approach, focusing on catching perpetrators rather than supporting victims and addressing harm [125, 126, 129, 193, 257]. These issues became more prominent with the widespread adoption of automated and AI-supported moderation systems [80, 235, 247]. Although muting and blocking allow players to take control in moderating toxic behavior [2, 235], they conceal the problem rather than resolve it, and fail to restore trust or promote community health [187].

A recent line of work explores how novel approaches, such as restorative justice mechanisms [257], real-time [76, 193], and AI-driven support [76, 216] can provide alternatives to the punitive and top-down moderation.

This dissertation contributes to this growing body of literature by investigating social safety, privacy, and moderation tools in online games. I show that current tools are insufficient for the broad player base and demonstrate the viability of leveraging supportive mechanisms to augment the overwhelmingly punitive apparatus.

Throughout this thesis, I focus on adolescents and young adults, who are a key demographic group. Gen Z users (ages 11 to 26) are especially vulnerable to toxic behavior; they go through substantial neurological, emotional, and social development while also extensively playing online. Adolescents make up the largest gamer demographic, accounting for 28%

of players in the USA [99]. They are also the most active social gamers, frequently playing online and engaging in esports, streaming, and adjacent online communities [51]. Online safety is particularly crucial for young players, as adolescence is an influential period of forming passionate commitments and social identities [22, 32, 49, 112], learning emotional regulation [84] and moral reasoning [56]. While the biological and emotional changes of puberty start in early adolescence [52], many aspects of neurocognition, including inhibitory control, problem-solving, and reflecting on choices, mature slowly and continue to develop into the mid-twenties [49, 97]. This developmental gap between biological and cognitive-emotional maturity creates vulnerability for teens, who navigate complex social situations and manage heightened emotions and peer pressure without a fully developed neural control system responsible for impulse control and evaluating consequences [10, 49]. Establishing suitable social and technological scaffolding that protects adolescents while supporting their growing need for autonomy, privacy, and respect is, therefore, imperative [49].

The thesis follows a three-study model, with the central research question of

What are promising future directions in moderation that support players in resisting toxic behavior and transforming toxic cultures?

This overarching question is explored through three studies investigating

1. young adults' response to toxicity and perception of moderation tools, and the relationship between the two,

2. the parental and player controls protecting players in popular online games and platforms, and

3. players' attitude toward real-time AI moderation and potential future interventions they envision.

The first two studies focus on current safety controls and moderation tools and how players

perceive them, while the third one explores potential future directions for using AI in real-time moderation, reflecting the players' perspectives. Together, these works contribute to a better understanding of current practices and systems, identify their strength and shortcomings, and point to largely unexplored areas of supportive moderation and real-time interventions that foster self-regulation in online games.

Starting with the players, in the first study, I explore how young adults respond to toxicity and perceive existing moderation tools. Both moderation and reacting to toxicity have a norm-setting effect. While moderation directly or indirectly nudges group norms [83], responding to toxicity can challenge, accept, or reinforce harmful conduct [11, 206]. Having a better understanding of how increased game experience and exposure to toxicity affect the responses to toxicity and the perceptions of moderation tools can help design better systems that support players in resisting toxicity rather than escalating harm. I also explore the alignment between players' habitual coping strategies and their endorsement of moderation tools to (a) see whether particular tools are more valuable for players depending on their coping styles, and (b) to identify which tools need improvement to fit a broader group of players. To explore these relationships, I use data from a cross-sectional survey of young adult players ($n = 602$) and conduct confirmatory factor analyses and structural equation modeling. My findings reveal that responding to toxicity is not role-based; players don't respond to harm differently as victims or bystanders. Instead, coping is a habitual strategy that also relates to the moderation tools they perceive as effective.

In the second study, I assess the full suite of parental and safety-related player controls, focusing on teens. Social safety and privacy are crucial aspects of their participation in multiplayer online games. To gain a comprehensive view, I conduct a systematic feature analysis of parental controls and safety tools in 12 popular AAA multiplayer online games and 6 accompanying publishing platforms. I use the Teens Online Safety Strategies (TOSS) framework [253] as a high-level organizing framework to explore how the offered safety mecha-

nisms align with different parental approaches (restriction, monitoring, and active mediation) and teen self-regulation strategies (self-monitoring, impulse control, and risk coping). For a deeper analysis, I borrow and adapt Wang et al. [241]'s extension of the TOSS framework and assess granularity: the level of control these tools provide to parents and teens, transparency: how clear the rules are for teens, and how clear the features are for the users, and the third axis, communication: how various features encourage communication between parents and teens. This work spotlights important shortcomings of parental and safety-related control systems. I identify where they leave teens vulnerable or ignore their need for autonomy, and pinpoint unexplored viable areas in the current safety and moderation framework.

Parental controls usually align with restrictive approaches, and it is often unclear how platform settings carry over to individual games. The controls of in-game safety, privacy, and communication are scattered over different menus, and the lack of granularity often leaves teens with limited options to customize their social experiences. These findings generalize to the larger player community.

After exploring players' perception of the moderation apparatus and the current state of safety controls offered in games, I turn to potential future directions. My findings of the first two studies highlight the need for a cohesive and transparent social safety infrastructure and the opportunity to complement existing tools with novel supportive mechanisms for self-regulation and active coping. The third, and final study, investigates such opportunities and invites players to creative ideation about real-time AI interventions they approve of and find helpful.

Through a reflexive thematic analysis of five focus groups with 28 young adults, I explore attitudes toward current AI moderation and potential alternative solutions. The results reinforce my previous findings on the need for supporting mechanisms and increased transparency, extending these with player control, agency, and customizability. Crucially, even with "the perfect AI" in mind, players reject many punitive mechanisms on ethical grounds

and favor de-escalation interventions. Integrating real-time support in the game world and leveraging players' emotional attachment to game characters show the highest acceptance and a promising new avenue for interventions.

1.3 Thesis Statements

- Proactive tools are few, and in-game supportive interventions are missing from the current moderation apparatus.
- Supportive mechanisms that promote self-regulation, including self-monitoring and real-time impulse control, seem particularly promising for teens and young adults. Unlike punishment, support can be optional, personalizable, contextual, and integrated into the game world; therefore, it can accommodate players' need for autonomy, control, and immersive play.
- The diegetic integration of moderation and leveraging players' emotional attachment to in-game characters is an auspicious avenue for nudging prosocial behavior, perspective taking, and self-reflection.

1.4 Thesis Structure

This thesis follows a three-article format for publication and is structured as follows.

In Chapter 2, I first define toxicity and its taxonomies, then introduce the literature on the contextual roots of toxic behavior. To give a broader context, I review the aspects of virtual spaces, gamer identity and belonging, and problematic design of game content, mechanics, and meta-systems that enable or foster harmful conduct. Finally, I formulate the problems around the shortcomings of technological moderation, the core subject of this thesis.

Chapter 3 examines how cumulative exposure to toxicity and game experience impact young

adults' responses to toxic behavior, their perceptions of in-game moderation tools, and potential relationships between the two. To model these relationships, I use confirmatory factor analysis and structural equation modeling on a cross-sectional survey of 602 adolescents and young adults. These results, as written in this chapter, are currently under review in ACM Games: Research and Practice.

In Chapter 4, I investigate the current parental controls and in-game safety, privacy, and communication settings that support teens' social safety in online games. For this effort, I adopt Wisniewski et al. [253]'s Teens Online Safety Strategies (TOSS) framework as an organizing framework and use its extension by Wang et al. [241] to conduct a systematic feature analysis of the safety-related parental and player controls in 12 popular multiplayer games and 6 publishing platforms. This chapter is prepared for publication in ACM Transactions on Computer-Human Interaction.

I explore player attitudes toward AI moderation in online games and the potential future directions in Chapter 5. Building on the results of the previous two chapters, here I focus on supportive real-time AI interventions that players identify as suitable and helpful in the games they play. I discuss the findings of a reflexive thematic analysis of five focus groups with 28 young adults. I intend to submit this chapter for publication as a conference paper in the ACM Conference on Human Factors in Computing (CHI).

Chapter 6 concludes the thesis with a summary of the key findings and contributions of my research. Revisiting my thesis statement, I discuss how the results of each study support the central argument for shifting from the punitive moderation paradigm to a supportive framework, which aligns better with player needs and shared expectations.

Chapter 2

Background

2.1 Toxic Player Behavior

2.1.1 Definitions and Taxonomies of Toxicity

Forms of toxic behavior in online games span across a broad spectrum, from benign mischief and contrary play to predatory and criminal conduct. However, there seems to be little consensus on defining toxicity and classifying behaviors that count as toxic due to its ambiguous nature and numerous interpretations across platforms, genres, communities, and cultures [24, 70, 114, 124, 137, 141, 175, 235]. Additional factors, such as the social cohesion among players, their moral position, and interpretation of the disruption, further complicate drawing a well-defined line between acceptable and unacceptable behavior [62]. To reduce this ambiguity, researchers have developed various taxonomies of toxic behavior (see, for example, griefing taxonomies [1, 65] and taxonomies of toxicity [124, 136]).

Toxicity is the most commonly adopted umbrella term for negative behavior, including both communicational and behavioral transgression [2, 207, 235] that adversely impact the experience of other players [82, 125, 130, 137]. Common forms of communication (verbal or

written) transgressions are spamming, trash talking, name-calling, flaming (offensive, hostile, derogatory, or profane comments often made personal), and inappropriate usernames. Behavioral transgressions include false reporting, inappropriate role-playing, contrary play, inhibiting the team/aiding the opponent (intentionally throwing the game, leaving, sabotaging), and cheating (using third-party tools, boosting, smurfing). Less frequent and more heinous toxic interactions are hate speech, threats of violence, doxxing, swatting, extremism, as well as criminal and predatory conduct.

Among others, cyberbullying [1], harassment and sexual harassment [2], trolling [3], and griefing [4] are the most frequently studied facets of toxic behavior, all of which assume the intention of harm or disruption. Poor player experience, however, can also result from unintended

[1]Similar to traditional bullying, **cyberbullying** compounds power imbalance, repetition, and intent to harm. In the case of cyberbullying, power imbalance results from differences in skills and ranking, and, although repetition does not happen between the same (cyber)bully and victim, it is in a way realized by sharing and circulating the harmful content in a larger audience [9]. Examples of cyberbullying include *"verbal and visual social and relational aggression, such as name-calling, rumors, flaming, harassment, denigration, sexting, outing, embarrassing photos or memes, stalking, and impersonation"* [9].

[2]**Online harassment** can be considered as a form of cyberbullying and defined as hostile or unwelcome behavior resulting in anxiety, distress, or discomfort for the target, regardless of the harasser's intentions [69, 70]. General harassment includes trash-talking, threatening, insulting, and hindering progress [9, 68, 69]. However, harassment often originates from discrimination. Forms of hate-based harassment target identities of gender, race, ethnicity, sexual orientation, religion, age, disability, or national origin. Accordingly, **sexual or gender harassment** is defined as unwelcome sexual attention and conduct that targets someone based on their sex, sexual orientation, or gender identity. Aside from sexual motives, discrimination against an "outgroup" – perceived as inferior – often leads to sexual harassment. In the context of online video games, it includes sexist jokes and insults, suggestive comments, unwanted sexual or romantic advances, as well as rape threats and jokes [9, 68, 69, 227].

[3]**Trolling**, another frequently analyzed aspect of toxic behavior, is also regarded as a type of cyberbullying that happens anonymously and seemingly without purpose [114]. A common theme in defining trolling is disruption, deception, and antagonizing other players, and that it is performed with hostile intent [37, 114, 235]. Trolling consists of insulting (offensive language or hate speech), spamming, and feeding (forfeiting the game by letting the opponents kill the player multiple times, 'feeding' [the enemy] is typical in team-based competition games, such as Multiplayer Online Battle Arena (MOBA) games), intentional fallacy, misdirection, inappropriate role-playing, griefing, flaming, and trash-talking [114]. This taxonomy significantly overlaps with the taxonomy of toxicity.

[4]**Grief play** or griefing typically involves breaking social rules and having fun or gaining advantage at the expense of other players while strategically trying to accomplish a goal, exploiting in-game mechanics [65, 175, 213]. Griefing can also be considered a game-specific trolling [142]. Foo and Koivisto [65] developed a taxonomy of griefing that includes 1) harassment (spamming, spatial intrusion, event disruption), 2) power imposition (repeatedly killing another player, e.g., rez-killing or newbie killing), 3) scamming (deceiving another player to gain advantage or in-game commodities), and 4) greed play behaviors. For example, taking resources or kills without earning them, e.g., camping and farming (monopolizing an area where the player repeatedly acquires resources, experiences, or points), ninja-looting or kill-stealing (taking experience points or resources that other players worked for and earned) [65].

disruption and tilt. **Tilt** refers to a negative emotional reaction to in-game events that ruins performance and decreases enjoyment. It is a cyclic process that spirals a player into toxicity [235] [5] In fact, the majority of incidents occur between players who joined the game to have fun and win, rather than to cause distress or harass others [163, 232]. For example, in competitive games, players often experience stress, anxiety, and powerlessness, which can decrease performance, cause team conflicts, and result in toxic behavior [124, 134, 235]. Mismatched expectations are another common reason for tilt and harmful conduct. One example is the previously mentioned griefing. Since the game mechanics afford grief play, players can view it as a pragmatic way to gain an advantage at the expense of others. However, when it is not an accepted/expected norm in the community, it can cause substantial distress and frustration [65, 175, 213]. Expectations can also differ regarding playstyles. Experimenting with different strategies and characters rather than playing the most advantageous so-called meta also frequently leads to tilting and toxic behavior in team-based competitive play [129].

Recently, Thriving in Games Group (TIGG) [6] pivoted from the term toxicity and employed **disruptive behavior** to describe any conduct jeopardizing player experience or community well-being. In their framework, disruptive behavior encompasses behaviors that do not align with the norms defined by the players and the community. The term captures that disruptions are not necessarily destructive and can arise, for example, from exploring new modes of play or having mismatched expectations. They define **harmful conduct** as a subset of disruptive behavior that causes significant emotional, mental, or even physical harm to players or others in their lives [62]. Similarly, Kowert [136] distinguishes dark participation and toxicity. and defines dark participation as *"any deviant action that takes place in games"* [136], and toxic behavior as the outcome of dark participation *" that causes harm to another player and can vary from culture to culture"* [136].

Highlighting its situational and dynamic nature, Kou [124] argues that toxicity should be

[5] Note that some research uses the same definition for toxicity. [115, 117, 118, 120, 122].
[6] Previously Fair Play Alliance.

understood as a process mediated by player perception and interpretation [124]. They claim that certain forms of toxic play emerge in the gap between static rules and dynamic norms, and remain undetectable by current rule-based moderation systems. Consequently, toxicity needs to be analyzed outside the boundaries of rules and rule-breaking [124].

This work builds on the definitions of Fair Play Alliance and Anti-Defamation League [62], Kowert [136], and Kou [124], considering **toxicity a situated course of actions and emotions that develops over time, disrupts player experience, causes harm to other players, and is subject to perception and interpretation.**

Throughout this thesis, I use harmful conduct, toxicity, and toxic behavior interchangeably. Other terms, such as cyberbullying, harassment, trolling, and griefing, will be adopted when a more nuanced specification is necessary or when a cited work purposefully makes these distinctions.

2.2 The Contextual Roots of Toxic Behavior in Online Games

Although negative personality traits, such as Machiavellianism, psychopathy, sadism [37, 142], and impulsivity (negative urgency and sensation seeking) [153] were found to positively correlate with harmful conduct, and demographic factors (age and gender in particular) may influence players' perception and perpetration of toxicity [91, 121, 137], **the context of interaction and the players' emotional states seem to explain toxic behaviors better than innate factors** [206].

Regarding emotional states, recent work explored the relationship between emotional self-regulation and toxic behavior [74, 255]. Impulse control was associated with being less inclined to tilt and even when tilted, more likely to reappraise the situation and plan behav-

ior, instead of venting emotions or stopping communication with teammates [255]. While emotion dysregulation has a positive effect on toxic behavior, suppressive emotion regulation (a controlled approach to minimize emotional inputs) is associated with less toxicity [74]. Self-regulation is a crucial facet of toxicity; it can escalate or resolve a conflict and can have a significant effect on building healthy coping strategies and improving overall culture. I'll discuss this issue further in the following chapters.

Next, I provide an overview of the contextual factors driving or enabling toxic behavior, including anonymity and online disinhibition, the facets of identity and belonging, and arriving at the games themselves: the architectures shaping play and participation, and the central topic of this dissertation: player safety and moderation systems. For the sake of completeness, I briefly introduce game characteristics, such as game content, game modes, and meta-systems, which are not central to this dissertation but belong to the broader concept of technological governance; they shape participation and behavior, and implicitly contribute to the problem of toxicity. Finally, I summarize the player safety and moderation tools commonly offered in online games. These systems collectively aim to create a safe and fair experience for the players. Therefore, instead of resorting to more common but somewhat limiting terms, such as intervention tools or moderation, I will refer to them as **safety infrastructure** throughout the dissertation.

2.2.1 Virtual Spaces and Anonymity

Contextual factors that emerge from the affordances of online play, such as anonymity and reduced nonverbal cues, are among the well-cited causes of toxicity [42, 69, 235].

The **online disinhibition effect** (ODE) [225] posits that anonymity grants individuals a sense of freedom to act without the social constraints of real-life interactions while maintaining privacy [9, 11, 117]. Toxic disinhibition was found to predict hostile behavior in online games [11, 119, 235] and is inversely related to the perception of toxicity's dangerous-

ness [11]. Similarly, according to the **social identity model of deindividuation effect** (SIDE), players in virtual environments perceive each other through limited social cues inferred from gamer tags, avatars, and computer-mediated forms of communication [147]. Due to this anonymity and limited identifiability, players experience deindividuation, loss of self-awareness, and reduced care for others [69, 70, 91, 206]. Furthermore, players may bypass their personal identities and behave according to the perceived social norms, even if those promote hostility and aggression toward others, especially toward individuals seen as "outsiders" [82, 226, 227]. In the context of online games, it means identifying with the salient social identity of the "gamer" [82, 226, 227]. Thus, gamer identity and belonging are other potential sources of harmful conduct.

2.2.2 Identities and Belonging

Gaming has a long-standing reputation as a male-dominated space with prevalent sexism, misogyny, and homophobia [9, 68], reinforcing masculine norms and the divide between members of the perceived in-group and out-group [4, 68]. Despite the increasing diversity of players [59, 60], **identity-based discrimination** is still widespread in games. Individuals who do not fit into the dominant heterosexual white young male identity are frequent targets of harassment [9, 11, 68, 69, 182, 213].

Players are often victimized based on demographic attributes, such as gender, race, ethnicity, national origin, sexual orientation, and the intersectionality of these identities [69, 182]. Being targeted for belonging to a marginalized group has distinctly severe consequences for the players [69].

Although computer-mediated communication in online games only reveals limited cues of the players' identities, it is often enough to make assumptions about gender, race, ethnicity, or national origin, and turn players into the target of derision and harassment [166, 173, 207]. An example of this is voice-activated racism, sexism, and nativist discrimination: an

interpersonal, aural stereotyping based on the speech patterns and cadence of a player's voice [173].

As Quinton "Rampage" Jackson, the 2007 UFC Light Heavyweight mixed martial arts champion, notes: "*when they came up with Halo 2 [Microsoft Game Studios, 2004], it was more like for the online thing. I stopped' cause my own team would kill me because they'd hear my voice and start calling me "nigger". "Nigger this, nigger that!" If you was right next to me at the arcade, you wouldn't say nothing. I don't ever tell them who I am, but I just forgive them for being ignorant.*" (quoted by Nakamura [173]).

Online game communities are the most inequitable regarding the treatment of women [69]. Female players receive disproportionate, and frequently sexual, attention when playing online [166]. Male gamers commonly act as gatekeepers, downplaying females' skills, questioning their credibility, and challenging them to prove themselves worthy of playing [166, 213]. Female players often experience being ignored and overlooked by men, refusing to listen to them or making them feel as if they need to be "carried" and couldn't perform well without men helping them [166, 213]. Additionally, females are often denied the true gamer identity through derogatory labels, such as "fake geek girls", who are considered to play video games with apparent sexual motives, or "casuals", who enjoy browser-based or mobile social games [227].

Aside from identity-based discrimination, but related to masculine norms, another facet of belonging is **skill level and expertise**. Time spent playing and level of game involvement were shown to predict general harassment [226] and cyber-aggression [91]. Prior research found that game rank is the most common reason for cyber-victimization [9] and that toxic behavior becomes more predominant in higher ranks [206]. Less skilled participants are often labeled as "noobs" [7] or "newbies" and are frequent targets of cyberbullying [42].

[7]Noob, in the context of multiplayer online games, refers to a player at a low skill level, who is perceived as someone neither willing nor capable of becoming better than a beginner.

When players with different skill levels are grouped in a competitive game, the power imbalance usually leads to toxicity: higher skill and experience levels create a stronger sense of belonging and a status advantage. This sense of superiority, together with a lack of perspective-taking and empathy, can trigger toxicity against less skilled, novice players, or those seen as outsiders [206, 266].

Identity, belonging, and in-group status shape participation and affect how players cope with toxic incidents. In Chapter 3, I will come back to these factors when analyzing the impact of toxic exposure and game experience on players' responses to toxicity.

Masculine norms, competitive play, ranks, and imbalanced team composition all strongly relate to the design of game mechanics, content, and meta-systems, such as progression, ranking, and matchmaking. The design and implementation of these game systems shape participation, social interaction, and the perception of belonging. Consequently, they may also enable toxic behavior and contribute to toxic cultures.

2.2.3 Game Design Enabling Toxicity

A long line of research argues that technological artifacts embody the values of their creators [161, 183, 249]. Designers inscribe "scripts" to technological artifacts that delegate responsibilities beyond functionality and guide user behavior [144, 237]. If technology mediates user behavior, then the design of these artifacts is inherently a moral activity [237] and the game architecture itself is a form of governmentality. Game developers are choice architects, designing and implementing game mechanics that influence player choices, such as forms of social interaction and the amount of time spent playing the game [111, 230].

2.2.3.1 Game characteristics

2.2.3.1.1 Game Content depicting **characters and their roles** in stereotypical ways perpetuates racial and gender bias and effectively socializes players into racism and misog-

yny [154, 155]. Sexist content, such as portraying female characters as weak or sexualized individuals, primes sexist attitudes and increases the likelihood of gender-based harassment [68, 140].

Examining the relationship between game content and culture-based toxicity, Şengün et al. [203] found that players recognize the real-world connotations of character designs, such as lore, weapons, outfits, and occasionally use them as a tool for targeted harassment of marginalized communities. They show that League of Legends champions, like Zilean, Kog'Maw, and Ziggs, are used to discriminate against players deemed Arab and/or Muslim. Because Zilean and Ziggs use bombs as weapons and Kog'Maw explodes after he dies, players associated them with suicide bombers [203].

Further, character design and customization have a significant influence on behavior. Deindividuated players not only conform to the social identity of their community, but they also embody the perceived identities of their avatars. According to the *Proteus effect*, players change their behavior based on their digital representations. A study about League of Legends champions [204] found that the level of toxicity and valence in in-game chat varied depending on the champion played. Selecting offensive roles, such as assassins, marksmen, and fighters, resulted in more toxicity and negative valence than tank and support roles. They observed similar outcomes for male champions, while the difficulty of the character had no impact on toxicity [204].

2.2.3.1.2 Competitive game mode is another well-cited trigger of toxic behavior. Studies found that players in ranked matches showed higher levels of toxicity [82] and reported experiencing more toxic exposure in more competitive games [2, 207]. Supporting these claims, Kou [124] found that players accept toxicity as an "inevitable byproduct of competitive gaming" and admit that this spurred them to flame their teammates [124]. Relating to competitiveness, stress and frustration have also been found to drive harmful

conduct. The more stress and frustration players experience as a team, the more toxic their behavior becomes. Acute stress is also negatively associated with the team's capacity for information processing, having a damaging effect on performance [206]. The emotional upheaval resulting from underperforming further ruins collaboration and performance, causes tilt, and results in spiraling to toxicity [235]. Additionally, Shen et al. [206] concludes that toxic behavior becomes more predominant as the player reaches higher ranks.

Games with team-based competition, such as first-person shooter (FPS) and multiplayer online battle arena (MOBA) games, often only offer **weak sociality**. In contrast to multiplayer online role-playing games (MMORPG), where groups are more persistent and long-term collaboration is essential for success, in competitive games, players join random teams and work for their individual goals as they climb the ranked system [124]. Even with social affordances, such as adding peers to a"friend" list, players typically do not foster relationships; they don't expect strangers to engage in socializing, and often intentionally keep distance to avoid the pressure to play [236]. When future interactions are unlikely once the match is over, self-interested players may deflect and put less effort into being friendly [206]. Further, due to the weak sociality, players instrumentalize teammates, judging them according to their rank and contribution, and seeing them as assets or obstacles in reaching their goals [124, 236]. Consequently, making mistakes or not performing as well as expected often results in flaming and toxic outbursts [124]. The ephemerality of matches and the lack of social ties also create an environment where toxicity is easily dismissed and has seemingly no consequences. As an Overwatch player notes: "*In games, you don't know these people. You don't know their personalities. You don't know their backstory* [...]. *All you know is that they're doing bad, and so that's what you use against them. And, there's no way of fixing that. And if you're getting offended by some of these words, then just mute them*" [257, p. 111:16].

2.2.3.2 Meta-systems and Neoliberal Values

Many multiplayer online games – MOBA in particular – have a subscription-based or free-to-play business model that relies on frequent participation and in-game micro-transactions. Online game design naturally has to cater to this need, but in the meantime, ethics, prosociality, and inclusion often get sidelined. Companies started to develop social architectures valorizing individual progress over civic-mindedness and building strong social ties [17, 45, 125]. Systems like matchmaking [45, 134], ranking [131], and rewarding [44, 111] embody neoliberal values, such as continuous competition, individual achievement, progression, and accumulation, in games that can never be completed. An example of this is how Blizzard has changed world events, the matchmaking system (Group Finder), and the phasing in WoW over time in a way that encourages players to play solo games and see each other "as obstacles of success" rather than allies and community [17, 44, 45]. A player describes how the changes promote greed play and disrupt sociality "*[...] open world combat was the stupidest thing they have ever done. In World of Warcraft, I am hunted down and killed repeatedly just because they can get a few points out of me for better items and gear*" [42, p. 8]. While others criticize the lack of meaningful social interactions: "*I used to play MMOs to play with other players and make new friends. These days it feels as if I was playing a single player game with some decent AI."* According to another player: *"Blizzard has killed basically all the social aspects of WoW...other players may as well be AI Bots for all you need to communicate with them.*" [46].

2.2.3.2.1 Progression systems with in-game purchases and unlockable items show a similar trend in multiplayer online games. Experience points are essential in leveling up to higher ranks and buying in-game artifacts. Rewarding and ranking mechanisms, therefore, are designed to keep players engaged in fast-paced, challenging, and sometimes frustrating play in pursuit of higher ranks [125, 226]. The game rewards spending more time playing, amassing resources, and winning, but it also effectively creates a rivalry between team

members. In League of Legends, for example, teammates compete for resources throughout the match to purchase items that strengthen their characters [61, 125]. In both DoTA 2 and League of Legends, those who assist in killing opponent NPCs or are within a certain perimeter are rewarded with in-game currency, but the player dealing the final hit gets the most sizable cut, which makes stealing kills worthwhile [61].

Losing, however, is costly in most competitive games. It decreases experience points and can ruin ranking. In many games, competition already starts before the match when players select characters and decide on a strategy. MOBA players, for example, who are deeply invested in progression, would go as far as to check their teammates' statistics on third-party websites and manipulate them into playing their top characters to optimize the chances of winning. As a League player noted: "*Go to lolking.net and open up a tab for every one of your 4 teammates. Click on "Ranked Stats," then sort by "Wins." Your new goal in life is to get your teammates to play one of their top 2 roles, you can judge by their wins with champs (if they also have a positive win rate). Force your teammates into their best role if at all possible*" [131, p. 7]. Even if cooperation, sportsmanship, and respect are expected of players, these are secondary to performance, achievement, and competition when the game mechanics imply that winning and rank have the highest value. In such environments, toxicity is often tolerated when the offender is a highly skilled player [125].

2.2.3.3 Ranking system

Rank has a more complex meaning to players than simply skill level; it is a significant part of their online identities. Games reward high ranks with in-game cosmetics, lending a distinct appearance to characters and items [33]. They consider it a form of social status with many dimensions, including knowledge, mastery, temper, and attitude. The lowest rank is associated with low achievement, reckless style, frequent raging, and unwillingness to learn and cooperate [131]. Players in this rank are often punished based on these stereotypes, even if they are simply new to the game, making noobs and newbies frequent targets

of harassment [9]. When a game mechanism conflicts with players' interests, they try to find workarounds and appropriate features to their needs. An example of this is creating secondary accounts to experiment with characters. Learning a new champion or role is a challenging task. Since matchmaking would happen based on previously established skill level (c.f. Sec. 2.2.3.4), the lack of experience and mastery playing the new character would result in underperforming and eventually degrade the rank, not to mention the frustration it could cause to team members. To avoid this issue, players often set up new accounts to practice and get skilled in playing a new champion or role [233]. Another example is related to how players define end goals for themselves. Multiplayer online games usually cannot be "finished": players continuously work on their progression, which becomes slower and more cumbersome as they climb the ranked system. Some set a goal for themselves in reaching a certain rank, and upon accomplishing it, they would be happy to stop. Stopping, however, is not aligned with the interests of game studios. Many games have decay mechanisms eroding player rank after a period of inactivity. In League, this only impacts ranks higher than Platinum. Kou et al. [131] found that League players who reached the desired goals strategically played just enough games to prevent decay and avoid the risk of sliding down a rank [131].

2.2.3.4 Matchmaking system

Aside from ranks and progression, another source of player grievance stems from the matchmaking system. The visible rank doesn't matter in matchmaking, which instead relies on players' matchmaking rate (MMR). MMR is a numerical score mostly calculated based on the traditional Elo rating system, – used in chess and other sports – a reliable and self-correcting method determining the relative skill level within the rating pool [33]. The exact formula of MMR calculation and how the two teams get selected are obscured from players to prevent them from abusing the system. In a study about League players, Kou et al. [134] explains how players left in the dark try to make sense of their experiences as a consequence

of the lack of transparency. Players engage in "theorycrafting" to find logical explanations for losing and winning streaks and optimize their strategies accordingly. They discuss their frustration about the system that guarantees a 50 percent win ratio for each player instead of each team. Therefore, according to their perception, it creates imbalanced matches to control individual win rates [134]. A player writes angrily: "*The matchmaking system forced a wonderful 50% win ratio to everyone regardless of their performance or the number of games played. (irony off). In a winning streak you win even if you just troll. In a losing streak you lose even if you do very well. This game can go to hell*" [134]. Further, the authors found that players were not only frustrated because of losing streaks, but also, knowing what comes next, they were equally unhappy and anxious during winning streaks [134]. Such an agitated mood can easily cause tilt and negatively impact several matches. In addition to issues related to the game's core mechanics, disruptive behavior also relates to the mechanisms of the safety infrastructure in games. To complete the overview of design impacts on toxic behavior, the next section summarizes the most common moderation systems.

2.2.4 Disciplining Toxicity and Encouraging Sportsmanship

2.2.4.1 Moderation in Online Games

Moderation in online games is similar to legal procedures; at its core, it focuses on identifying and disciplining offenders [125, 193] and follows the subsequent steps of regulation, reporting, adjudication, and punishment [130].

Game companies use legal documents to govern player behavior on their platforms, including terms of service, privacy policies, and end-user license agreements. These contracts are usually obscure, hard-to-read legal texts [23, 238], with broad statements and ambiguous boundaries of acceptable behavior that cannot serve as meaningful guides of expected player behavior but rather aim to protect companies' interests and intellectual properties [23]. In contrast with binding contracts, codes of conduct and community guidelines usually have

a more legible style but are often inaccessible otherwise. Codes of conduct often resemble policies of misconduct rather than defining shared values and expectations [81]. Additionally, policy documents commonly place the responsibility of maintaining the norms on the players [23, 81]. A tool assigned to this responsibility is the reporting system.

Reporting, the entryway to the moderation cycle, and the only step where players are involved, is the most common mechanism in online games. After filing a report, toxic incidents are adjudicated, and depending on the frequency and severity of the offense, the perpetrator gets a penalty ranging from warnings through content removal, to the suspension of the player's account [125, 130, 257]. Such a disciplinary approach is typical in online games despite its well-known shortcomings, such as drawing the perpetrator's attention to punishments rather than the harm they caused and fully neglecting the victim's needs [257]. Due to the lack of transparency in decision-making, the reporting system is one of the least trusted and most misused tools. Players usually either refrain from reporting offenders [2, 11] or they doubt the efficacy and fairness of the decision making process [129, 137, 257]. Additionally, suspending the player's account proves generally futile: players learn to evade it by maintaining multiple accounts or keeping toxic conduct 'subtle enough' [124, 125, 257]. As a striking contradiction, a deleted account wipes all traces of harmful conduct, granting toxic players a clean slate [125]. Muting and blocking are the most favored instruments; they provide agency and immediate results, hiding verbal aggression and helping avoid future interactions [2, 213, 235]. Despite helping players shield themselves, withdrawal cannot address harm, restore trust, or foster thriving communities [187].

2.2.4.2 Encouraging Prosocial Behavior

Other efforts expand moderation with proactive prevention, such as priming and reputation systems. In 2012, Riot Games designed priming to stimulate positive behavior. They experimented with exposing players to either encouraging or discouraging messages in three different colors before matches [82, 163, 235]. They found that warnings about harassment,

verbal abuse, and offensive language reduced toxic interactions[163]. The result, however, was too temporary and subtle to bring cultural change [163].

The reputation system is another tool in the disciplinary apparatus that helps enforce norms and encourage prosocial behavior [233]. More and more popular online games such as League of Legends (Honoring), Overwatch 2 (Endorsement), and CS:GO (Commending) implemented such systems enabling players to acknowledge friendly, supportive, and sportsmanlike behaviors after each match [2, 61, 82]. The system manages a reputation that follows the player as a badge signaling their character [82]. Commended players are rewarded with experience points and in-game loot for their continuous good attitude. Both Overwatch and League have five levels, with the highest being the most reputable, and both systems use zero to mark dishonorable participants. Players can only provide positive recognition, but the disciplinary system sets the level to zero when a player gets a chat restriction or temporary ban[8]. A decay mechanism – similar to rank decline – also demotes the reputation if a player doesn't get endorsed for a while [233].

As seen, different facets of technological governance, from game architectures and content design to the safety and moderation apparatus, all shape participation and player behavior. But while genres and games themselves are often too different to yield generalizable and comparable results, safety infrastructures and the tools of technological moderation are consistent across titles, affording a natural comparison.

In the following chapters, first, I explore how players respond to toxic behavior, how they perceive the effectiveness of moderation systems, and how these two relate. Then, I analyze the safety instruments across popular multiplayer games and distribution platforms. Because many popular games are targeted toward teens and adolescents or played by them despite the mature age ratings, I extended the analysis to include parental control systems as well. Finally, I explore potential future directions and player perspectives on using real-time AI-

[8]League of Legends Honoring and Overwatch Endorsement

supported interventions to curb toxic behavior.

Chapter 3

Leave, Mute, or Retaliate? How Young Adults Handle Toxicity in Multiplayer Online Games

Multiplayer online games are a prominent entertainment medium with diverse player demographics. Research shows that people of all ages, genders, and ethnicities have taken up gaming as a leisure activity [58, 59]. Despite well-known benefits of online gameplay (e.g., [77, 86, 87, 174, 219, 222, 265]), toxicity between players remains its chronic darker side, adversely impacting players [48, 243, 254, 268], team dynamics [2, 175, 206], and the larger gaming community [11, 73, 91].

Toxicity is a widely adopted umbrella term encompassing a broad range of verbal and behavioral transgressions [2, 207, 235] that damage player experience [82, 125, 130, 136, 137]. Examples of toxic behavior comprise harassment, griefing, trolling, cyberbullying [2, 233, 266], incivility, anti-social behaviors, verbal aggression, degrading comments [206], mischief, cheating [2, 114], and, among the more egregious forms: hate-based harassment, extremism, and

criminal or predatory conduct [62]. In their Disruption and Harms in Online Gaming Framework [62], the Thriving in Games Groups (TIGG, formerly Fair Play Alliance) shift away from using the term toxicity in favor of disruptive behavior (conduct that does not align with community norms) and harmful conduct (a subset of disruptive behavior encompassing transgressions that cause emotional, mental, or even physical harm to player) [62]. Previous research has also highlighted the ambiguous boundaries of toxicity and its dynamic, subjective, and situational nature [62, 120, 124]. In this chapter, I retain the term toxicity for its broad use and ready understanding among developers and players, focusing on verbal and behavioral transgressions that, regardless of the perpetrator's intention, damage player experience and community well-being [124].

Although the toxicity found in online games is part of broader trends in online and offline social spaces, there is still much that designers can do to foster healthy gaming communities. The design and implementation of moderation apparatus significantly impact player interaction, experience, and the perception of rules and consequences. Understanding how players respond to toxicity, their attitudes toward the current moderation tools, and how the two may relate is crucial for improving safety and trust in multiplayer online games. With this work, I aim to contribute to this understanding.

In this chapter, I focus on adolescent and young adult players, a key demographic group that plays online games extensively while also in the midst of substantial neurological, emotional, and social development. Gen Z users (aged 11 to 26 years) make up the largest gamer demographic, accounting for 28% of players in the USA [99]. They are also the most active social gamers, frequently engaging in multiplayer online games, esports, streaming, and games-related online communities [51]. Safety is particularly vital in the case of young players, as this susceptible and influential period is crucial in identity formation [112], emotional regulation [84], and moral reasoning [56].

I apply a cross-sectional survey and structural equation modeling to investigate how young

adults' responses to toxicity and their attitudes toward in-game moderation tools vary by game experience and by exposure to toxic behavior as a victim and bystander. I then explore potential relationships between players' customary responses to toxicity and their attitudes toward the tools afforded by the industry.

This chapter aims to contribute to online games research and development in the following three ways. First, this work advances the existing research base on player coping strategies by exploring potential differences between bystander and victim responses to toxicity and the effect of gameplay experience and toxic exposure on response strategies. Previous work has discussed that the rapid transition between bystander, victim, and perpetrator roles during play creates a cycle of harm [38, 120], and examined potential coping differences in the context of these roles [37]. Moreover, while research indicates that frequently targeted, vulnerable players cope differently than non-vulnerable players [41, 68, 69, 166, 173, 182, 254], no studies to date have examined differences based on gameplay experience, although belonging and skill are well-established factors in who is targeted for abuse [206, 207]. My work aims to fill these gaps.

Second, this work contributes to research on in-game moderation tools by investigating possible relationships between player perceptions of tools and other key player variables. Research suggests that coping behaviors depend in part on the intervention tools made available [72, 193], yet few empirical studies (e.g., [72]) examine potential relationships between the two with any specificity. Here, I connect perceptions of intervention tools with prior bystander and victim exposure and gameplay experience, as well as with one's habitual responses to toxicity or "coping strategies".

Finally, by considering a wider set of games and genres, this data more broadly generalizes previous findings on attitudes and perceptions of in-game player moderation tools studied in the case of individual games [125, 129, 159, 160, 233].

3.1 Background

3.1.1 Coping Strategies and Responses to Toxicity

The extant body of research on player responses to toxicity is both deep and broad in terms of focus and definition. Prior work examines the experiences and coping strategies against toxicity generally [72, 182, 193, 213] and also specifically in esports [2, 12, 93, 152, 188, 212, 235], virtual worlds [31], and in the contexts of trolling [37, 39] and female player experiences [41, 68, 69, 166, 254].

To understand player responses to toxicity, previous research either employed existing theoretical frameworks of other domains, such as [146]'s transactional model of stress [93, 152] or, more commonly, classification models constructed on the basis of individual data corpora. For example, Lazarus and Folkman [145] introduced a response categorization distinguishing *emotion-focused coping* and *problem-focused coping*. The former attempts to reduce emotional distress and regulate emotional responses without confrontation, while the latter involves actions to solve the problem or minimize its effects [145]. Subsequent studies of coping with stress in esports [12, 188] and toxicity in multiplayer online games [72] adopt an extended model that includes *avoidance coping*, which seeks to disengage from the situation, including both psychological and behavioral distancing. Smith et al. [212] added two further categories by separating *approach coping* (long-term strategies to reduce stressors, such as planning) from problem-focused coping and including *appraisal coping* (cognitive re-evaluation of a situation to reduce its impact) [212]. The majority of prior work, however, developed taxonomies of responses to toxicity based on single data corpora. Studies taking a qualitative approach use thematic analysis [37, 39, 93, 152, 166, 182] or grounded theory [41] to derive a variety of high-level categories, while previous quantitative work employs exploratory factor analysis to uncover underlying structures and potential classifications of coping [69, 72].

Despite the diverse categorization of response types, research has found similar tendencies across different contexts and player groups. While players often view problem- and emotion-focused coping as more effective, they tend to rely more on avoidance strategies [12, 212]. Players most commonly mute, block, and ignore perpetrators or accept toxicity as a normal part of online play [12, 39, 72, 188, 212, 235]. Diffusing the situation and confronting perpetrators [72, 235] or switching between contexts (game mode, team) [72, 182] are less frequent strategies, and support-seeking is the least common approach [39, 72, 182].

While prior work includes response strategies that players develop over time (e.g., acceptance) or apply only outside of the game (e.g., talking to friends, seeking online support, watching gameplay tutorials) [72, 152, 152, 182], here, I will consider immediate in-game responses. Instead of exploring how players deal with toxicity holistically or whether the employed strategies effectively help them feel better, my goal is to analyze responses to toxicity based on their impact on the community and its norms. Players' reactions to toxic situations have a profound impact on norm formation and community well-being. When players stay silent or respond with equal harm, they contribute to the normalization of toxic culture [11, 37]. Therefore, I organized coping dimensions into four categories: (i) *productive*, signaling that the behavior is unacceptable, ii) *social withdrawal* and (iii) *game withdrawal*, both shielding players from harm without challenging the toxic deed, and (iv) *escalation*, further encouraging and reinforcing toxicity.

One factor with a possible impact on how players respond is whether they are the direct target of toxicity ("victim") or are witnessing the altercation occur ("bystander"). It is less established whether player responses to toxicity vary based on their position as a victim or a bystander. Cook et al. [37]'s study of online trolling (the instrumental exploitation of game, chat, or website mechanics, at another person's expense [35]) indeed found differences. While victims often rage, ignore, or reciprocate toxic behavior, bystanders tend to join in on the conversation without lashing out or trolling in return. Similarly, studying MOBA games,

Kordyaka et al. [120] observed that although the bystander, victim, and perpetrator roles are fluid and players often transition between them even during a single match, bystanders have the capacity to support victims, while victims most often retaliate [120]. Quantitative research on how bystander and victim positions and exposure to toxicity as bystander and as victim impact responses to toxic behavior, however, is currently scarce. Therefore, I focus on the following research question:

> ***RQ1*** *Are there differences in how young adults as bystanders and as victims respond to toxicity in multiplayer online games?*
>
> > ***RQ1a*** *Does the frequency of bystander and victim exposure impact responses to toxic incidents in multiplayer online games?*

Not all players face toxicity to the same degree, however, and their status in the game contributes to varying rates of in-game toxic exposure. Passmore and Mandryk [182] found that responses to toxicity depend on intersectional identities and systemic inequities, ranging from the privileged few, who can afford to be skeptical of discrimination and tend to cope with the "nuisance" of toxicity through ignoring or enduring it; across the emboldened many, who adopt problem-focused responses strategies, fighting back until losing hope and retreating; to the conflict weary, typically with intersecting marginalized identities, who anticipate discrimination and tend to use emotion-focused or intrapersonal coping [182]. Following Allport [4], I use the term "in-group" to refer to players who occupy established, high-status, or majority positions within the community, and "out-group" to describe players who are newer, lower-status, or otherwise marginalized within that same social context. When dealing with toxicity, "in-group" players often start with low-effort strategies; most often, they ignore, avoid, or accept toxic behavior, and only seldom attempt to diffuse the conflict, switch game or context, or seek support [2, 12, 72, 212, 235]. "Out-group" players, in contrast, commonly use avoidance strategies such as hiding their identity, which may itself inadvertently mark them as targets, withdrawing from voice communication, or even the game

entirely [41, 68, 69, 166, 213]. For example, the experiences of female players with toxicity are well-documented. In addition to the above-mentioned strategies, women also frequently hide their gender to protect themselves [41, 69, 166, 213, 235], refraining from voice communication [41, 213], using gender-neutral gamer tags, and foregoing feminine avatars [69]. All despite the significant disadvantage of such responses in team-based competitive games [257] and the fact that it renders female players invisible, inadvertently contributing to the perception that female players are rare, if not entirely absent [68]. Other "out-groups" frequently targeted by toxicity are novice and unskilled players [70, 207]. Teams with significant skill disparities among their members are more likely to breed toxic behaviors [206]. Higher skill and experience levels bestow perceived and real advantages in games. A sense of superiority combined with a lack of perspective-taking and empathy can trigger toxicity against less adept players [206, 266]. However, while responses to identity-based discrimination are well documented, little research exists exploring variations in responses based on experience and perceived skill. Therefore, I consider the subsequent research question:

> ***RQ1b*** *How does gameplay experience influence responses to toxicity in multiplayer online games?*

3.1.2 Moderation in Multiplayer Online Games

To protect their communities, online platforms, including multiplayer online games, practice custodianship by moderating participation [78]. *Moderation is a set of governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse* [83] p. 52). Such mechanisms include platform policies and community norms that prescribe expected behavior, as well as technological systems that enforce them, such as constraints embedded in the game code and player-facing moderation tools [83, 125]. Here, I focus on tools that are relatively similar across games and platforms and allow comparison, including policies and moderation tools developed to curb disruptive behavior. Unfortu-

nately, most of these tools focus on identifying and punishing offenders [125] rather than supporting player agency and empowerment [193], and only a few are designed with prevention in mind [246, 247].

3.1.2.1 Reactive Moderation

3.1.2.1.1 The Core Moderation Mechanism – Punitive Approach. Networked games implement complex moderation apparatuses to mediate play and participation. Similar to legal procedures, these systems follow a coordinated sequence of regulation, reporting, adjudication, and punishment [130]. *Reporting*, the only step involving players, is one of the least popular mechanisms. Players are excluded from adjudication and seldom receive feedback regarding their reports, eroding trust in the efficacy of the system [129, 257] and leading to reluctance to use it [2, 11, 127]. As an outcome of reporting and adjudication, the convicted player receives a penalty based on the frequency and severity of the offense, ranging from warnings and content removal to communication restrictions and account suspension [125, 130, 257]. Despite their known issues, such punitive systems remain widespread.

Banning, suspending a player's account temporarily or permanently to prevent any participation, is the most common (albeit insufficient) penalty. Players view banning as a calculated risk and develop strategies to evade it, such as maintaining multiple accounts [125, 257] or committing subtle enough toxicity to bypass moderation systems entirely [124].

3.1.2.1.2 Complementary Moderation Tools – Withdrawal Approach. By contrast, instruments, such as muting and blocking, which provide control to the players to moderate toxic incidents themselves, and have an immediate effect, are the most popular [235]. Muting hides the selected player in voice and/or text chat, while blocking prevents friend requests and future communication. Both successfully shield players from harm by instantly hiding verbal aggression and restraining future interactions [2, 213, 235]. However, while such tools do provide players agency, they are still reactive, do not address the harm

done, nor restore trust and safety after a breach of conduct, or otherwise help foster healthy communities [187].

3.1.2.2 Preventive Systems

3.1.2.2.1 Priming & Reputation Mechanism – Proactive approach Recent proactive efforts, such as priming and reputation systems, attempt to move beyond traditional punitive moderation and instead nudge players toward prosocial behavior and positively reinforce sportsmanship. Priming aims to stimulate positive behavior by targeting players with encouraging or discouraging messages right before matches [82, 163, 235]. The priming system implemented by Riot Games is a case in point, with temporary and subtle effects on dissuasion from harassment, verbal abuse, and offensive language [163]. Reputation systems are another tool designed to nudge players toward greater adherence to social norms by enabling players to formally recognize and commend friendly behavior and sportsmanship [2, 61, 82, 233]; examples include honoring in League of Legends (2009), endorsements in Overwatch (2016), and commending in Counter-Strike: Global Offensive (2012).

Researches across all three types of moderation systems (punitive, withdrawal, and proactive) although highlight the shortcomings of each, have important limitations: the bulk of these studies focuses on a narrow set of games – most commonly, League of Legends (2009) and Overwatch (2016) – and a narrow set of tools, specifically, reporting tools [129], permanent banning [125, 159, 160], and reputation systems [233](cf. [257, 258]). Research to date is largely qualitative (cf. [159]), prohibiting comparison of players' perceptions of the efficacy of the broad suite of common moderation tools along a single consistent scale. Finally, no prior work examines how bystander and victim exposure to toxicity and gameplay experience impact tool assessment.

In this work, I expand the set of included games and collect quantitative data about the full suite of commonly available moderation tools to obtain a broader understanding of players'

perceptions of their comparative efficacy and generalize prior findings. I connect perceptions of moderation tools with prior bystander and victim exposure and gameplay experience. Toward these ends, I inquire

> ***RQ2*** *How do young adults perceive the efficacy of moderation tools and mechanisms?*
>
> > ***RQ2a*** *Do these perceptions change with the level of gameplay experience?*
> >
> > ***RQ2b*** *Does the frequency of bystander and victim experience impact these perceptions?*

Research suggests that player responses to toxicity depend in part on the moderation tools made available; yet, we know little about the potential relationships between the two with any specificity. A previous survey study by Frommel and Mandryk [72] found that in-game access to moderation tools and the frequency with which they are used directly affect how players perceive the effectiveness of related coping approaches. I broaden this work by exploring possible connections between players' perceptions of tool efficacy and players' preferred response strategies to toxicity, as reflected in my third and final research question

> ***RQ3*** *Are there any associations between responses to toxicity and perceptions of moderation tools' effectiveness?*

3.2 Methods

3.2.1 Instrument

The data included in this paper were part of a larger survey study focused on toxicity, hate, and extremism in online games, the experiences young adult players had as bystanders, victims, and perpetrators of such behaviors, their attitudes toward the problem and toward moderation tools provided by industry, and the possible normalization of such behaviors.

Here, I present my findings from three specific subsections of the full survey instrument: (1) players' gameplay experience, (2) experiences of and responses to toxicity as victims and bystanders, and (3) perceptions of the efficacy of player moderation tools. (See the Appendix A for the survey items.)

Demographic items were designed in accordance with current best practices for inclusive survey design [96, 108, 167]. Participants reported their age (in years), gender (female, male, nonbinary, prefer not to answer, or prefer to self-describe), and ethnicity (Asian, Black/African, White/Caucasian, Latinx/Hispanic, Native American, Pacific Islander, or prefer not to answer). Item phrasing and response options followed inclusive data-collection guidance encouraging flexible identity categories and the inclusion of self-description and opt-out choices.

To measure participants' **game experience**, I included behavioral and identity items grounded in prior gaming literature. I asked about the *amount of gameplay* (hours per week) and *duration of gameplay* (years playing online) as standard usage metrics commonly employed in game behavior research (e.g., in [16]). I measured *gamer identity* ("I identify as a gamer") using established scales, such as the Gamer Identity Scale [263], which operationalizes gamer self-concept. Furthermore, I assessed the frequency of *competitive play* and the *perceived expertise / self-assessed skill* ("I am good at videogames"; "I am skilled at the games I play"), analogous to items in validated gaming skills questionnaires (e.g., Gaming Skills Questionnaire [267]). Finally, genre preferences were solicited via a "select all that apply" list of common genres plus an open "Other" text field, following standard practice in player surveys to capture multidimensional play profiles.

Participants' **exposure to toxicity** was measured using nine items, designed to capture a graded continuum of online harms, ranging from lower-intensity social friction to high-severity abuse. Items included *embarrassment, offensive name-calling, trolling or griefing, harassment, hate-based harassment* (harassment targeting a person based on their mem-

bership in a demographic minority), *threats, doxxing* (making public personally identifiable information about an individual without their consent), *stalking* (unwanted and/or repeated surveillance or contact by an individual or group toward another person), and *swatting* (criminal harassment that weaponizes emergency services by them to another person's location without cause).

These categories were adapted from large-scale analyses of online harassment [57, 148] and research on toxic gamer culture and online aggression [34, 118, 141, 199]. All 9 items were rated on a Likert-type frequency scale and presented as three separate matrices for three participant roles: bystander, victim, and perpetrator. This tripartite framing draws on participant-role approaches to aggression [198] and extensions to digital and gaming contexts [69, 240], reflecting how individuals may move fluidly between observing, experiencing, and enacting toxicity online [38, 120].

Participants' **responses to toxicity** were measured using items representing four empirically grounded categories: *productive*, *social* and *game withdrawal*, and *escalating* responses, which capture distinct ways players react to toxic behavior in multiplayer contexts. This categorization builds on coping theory [25, 145] and its application to online gaming, where *problem-focused*, *emotion-focused*, and *avoidance* coping are distinguished [6, 12, 72, 188, 212]. *Productive responses* – including *reporting* a perpetrator, *refocusing chat* on gameplay, *supporting or seeking support* for a victim, and *calling out toxic behavior* – reflect *problem-* or *approach-focused coping* that actively signals a norm boundary and attempts to mitigate toxicity within the community. These strategies parallel confrontation, diffusion, and social-support behaviors identified in prior work as less common but more prosocial responses to toxicity [72, 235]. Social and game withdrawal represent two distinct levels of avoidance or emotion-focused coping, which protect the player but leave the toxic act unaddressed. *Social withdrawal* includes *ignoring, muting, and withdrawing from chat*, while *Game withdrawal* encompasses *leaving the match, taking a break, or quitting indefinitely. Escalating* includes

laughing along or *responding with toxicity in return.* Both withdrawal and escalation map onto maladaptive or retaliatory patterns described in studies of trolling and online aggression [37, 39], where bystanders sometimes normalize or amplify toxicity instead of challenging it. This structure reflects the broad consensus in prior research that, while players often view problem-focused coping as more effective, they tend to rely more heavily on avoidance and normalization of toxicity [12, 72, 188, 212, 235].

To measure players' **attitudes toward the moderation tools**, participants rated the effectiveness of 12 common moderation tools on a 5-point Likert scale ranging from (1) do not exist and (2) ineffective to (5) very effective. Participants who selected (1) *the tool does not exist* were removed from subsequent analysis. Tools were grouped into three categories. *Proactive tools* attempt to moderate behavior before toxicity begins, such as policies, priming, and honoring. *Protective tools* allow players to safeguard themselves against harassment and abuse, for example, muting and blocking. *Punitive tools*, like banning and reporting, to discipline or penalize perpetrators (cf. [2, 72, 182]).

The survey instrument was piloted twice, with 268 and 157 respondents, respectively, to ensure that item wordings are accurate and unambiguous. Items covering sensitive topics (e.g., actions as a perpetrator of toxicity) were preempted with text, assuring that all responses would be anonymized.

3.2.2 Procedures

This study was approved by the host institution's Institutional Review Board. Self-reported player data was collected using an online survey instrument via Qualtrics [190]. Participants were recruited through the social media networks of the Higher Education Video Game Alliance (HEVGA) and the North America Scholastic Esports Federation (NASEF) and from the undergraduate classes at the host university. Eligible participants were between the ages of 13 and 25 with prior experience playing multiplayer online games. After completing the

pre-screening form, those who met the criteria received detailed study information and a consent form (assent form in the case of minors, accompanied by parental permission and consent form). Written informed consent was obtained from all individuals, and in the case of minors, from legal guardians/next of kin. Upon consent, participants accessed the online survey, which took approximately 30 minutes to complete. Each participant received a 10 USD Amazon gift card for their contribution. After removing unreliable and incomplete responses, 602 participants remained, producing a 95% confidence level and a ± 4% error margin. Data were cleaned and anonymized on a private server. The data of this study are available on request.

3.2.3 Data Analysis

First, descriptive analysis was conducted to summarize the sample data [217] (See Appendix B for a summary of descriptive statistics), followed by exploratory factor analysis (EFA) to examine the factor structure of bystander and victim exposure variables and confirmatory factor analyses (CFA) to validate the theorized factor structure of player responses to toxic behavior and perception of in-game moderation tools. After assessing the validity and reliability of the measurement models, two structural equation models (SEM) were used to test the effects of game experience and toxic exposure on (1) coping strategies and (2) the perception of in-game tools. Demographic control variables, age, gender, and ethnicity, were added to both models. Finally, a third confirmatory factor analysis investigated the relationships between coping factors and tool perceptions.

Statistical analyses were performed using the R framework 2026.01.0+392 [191] with lavaan 0.6-21 [195], lavaangui 0.3.2 [106], semTools 0.5-7 [103], and dplyr 1.1.4 [245] packages. The exogenous categorical variables, gender and ethnicity, were coded with “male” and “caucasian” as reference categories, respectively. To estimate model parameters with ordinal endogenous variables, the WLSMV estimator was applied, which uses diagonally weighted least

squares (DWLS) and the full weight matrix to compute robust standard errors [195, 260]. Due to its sensitivity to sample size and model complexity [95], the χ^2 statistic is reported but not used as the sole basis for accepting or rejecting the models. Instead, additional fit indices were utilized to evaluate model fit, such as the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), the Standardised Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). I define cutoff values as follows: CFI and TLI close to $>$ .95 good and $>$ .90 acceptable, SRMR close to $<$ .08 good, and RMSEA close to $<$ .06 good and $<$ .08 acceptable [20, 95].

3.2.4 Sample Demographics and Game Habits

In this investigation, I adopt the expanded definition and period of adolescence that includes both teens and young adults between 13 and 25 years of age, a period defined by the neural changes occurring during this developmental window, the earlier onset of biological changes marked by puberty, and the sociological shifts in delaying responsibilities that mark adulthood into the early 20s [97]. Study participants (n = 602) were recruited primarily from universities and high schools in the coastal southwest of the United States. Their age ranged from 13 to 25, with an average of 19.87 years (SD= 1.77). Because most of the participants were 18 or older (97%), I use the term young adults instead of adolescents throughout this work.

Participants were primarily of Asian descent (60.1%), followed by Caucasian (18.3%) and Hispanic (12.0%) ethnicities. Most of them identified as male (66.1%) and as heterosexual (74.6%). 4% reported having a disability. Most participants were college students (63.3%), college graduates (18.3%), or enrolled in high school (14.5%).

The majority of participants reported playing online games between 5 and 20 hours per week and have been gaming for 8 - 12 years. The most preferred online game genres were multiplayer online battle arena (MOBA), sandbox, and shooter games (see Figure 3.1), which

are genres known to be most appealing to this age group [51] and feature both ranked game modes and team-based competition, typically associated with increased toxicity [2, 82, 207]. Participants predominantly identified as gamers (72.6%) or described themselves as at least moderately expert in the games they play (87.9%).

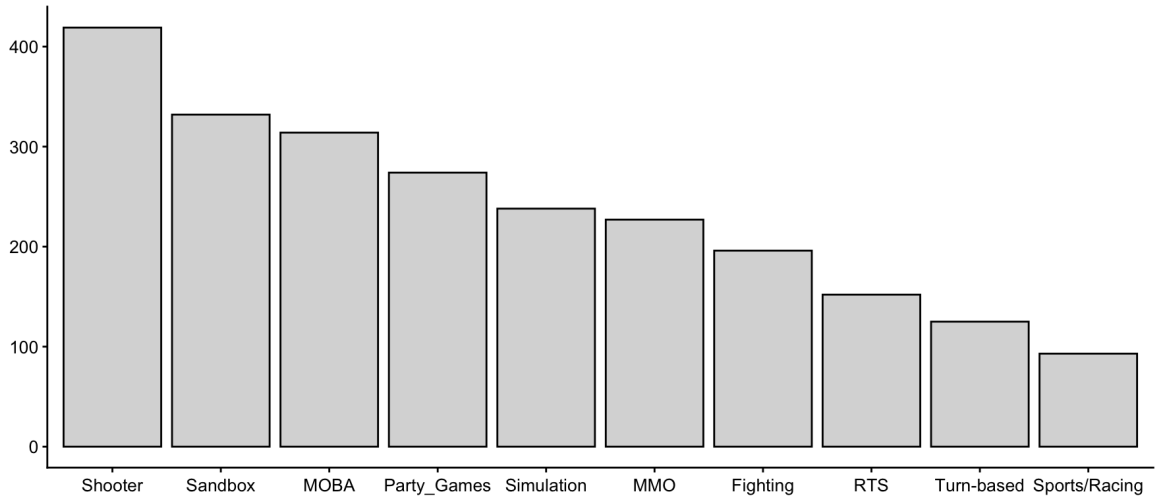

Figure 3.1: Preferred Video Game Genres

3.3 Results

3.3.1 Preparing Indicator Variables

In the first two groups of research questions, I use bystander and victim exposure, as well as game experience, as the predictors of responding to toxicity and perceiving moderation tools.

The data on exposure history reveal that nearly all (98.3%, $m = 2.1$, $SD = 0.9$) participants reported witnessing or having been the target (93.9%, $m = 1.5$, $SD = 0.6$) of toxicity in online games. Figure 3.2 shows the relative frequency of experience within each role by type of toxicity. While bystanders perceived that purposeful embarrassment and using offensive names were the most common types of toxic behavior, victims reported trolling/griefing as the most frequent item.

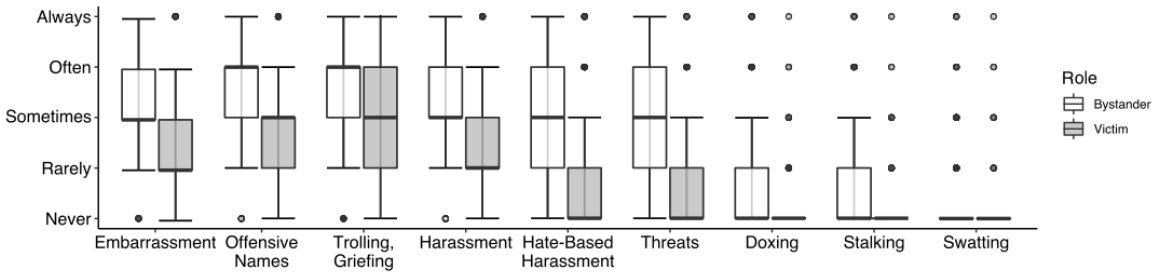

Figure 3.2: Bystander and Victim Experiences of Toxicity by Type.

Next, I performed an exploratory factor analysis (EFA) for bystander and victim exposure to examine the dimensionality of the observed variables. Both bystander and victim exposure were measured on 9 factors: name-calling, griefing, harassment, embarrassment, threatening, identity-based harassment, stalking, doxxing, and swatting. Kaiser-Meyer-Olkin (KMO) values and Bartlett's test of sphericity showed that the data are suitable for factor analysis, with an overall measure of sampling adequacy of 0.87 for bystander and 0.86 for victim exposure. Eigenvalues and scree plots indicated two factors for both sets of items. The weighted least squares factoring method was used with varimax rotation, allowing factors to correlate. Item loadings for both groups were strong and revealed the same structure for bystander and victim groups, with the first 6 items loading on the first factor and the last 3 on the second. The two factors explained 61% of the total variance for both bystander and victim exposure. There were no cross-loadings, and the groups were clearly distinguishable. The EFA is summarized in Table 3.1. The first construct represents more common forms of toxic behavior, including name-calling, embarrassment, griefing, harassment, identity-based harassment, and threatening, while the second reflects high-severity but less frequent forms of harm: stalking, doxxing, and swatting. To keep the models more parsimonious, in my subsequent analyses, I will use only the first construct representing bystander and victim exposure to common forms of toxicity.

In addition to the two derived latent constructs (bystander exposure and victim exposure), two composite measures, *game time* and *game experience*, were created by summing the z-

	Bystander exposure			Victim exposure		
Factor	1	2	h^2	1	2	h^2
Name Calling	0.88		0.70	0.84		0.71
Griefing	0.83		0.60	0.69		0.48
Harassment	0.81		0.65	0.86		0.76
Embarrassment	0.72		0.53	0.66		0.52
Threatening	0.64		0.53	0.68		0.52
Identity-Based Harassment	0.62		0.52	0.64		0.53
Stalking		0.83	0.66		0.83	0.30
Doxxing		0.76	0.66		0.77	0.36
Swatting		0.82	0.61		0.85	0.27

Table 3.1: Factor loadings and communalities (h2) of bystander and victim exposure latent constructs

scores of the related observed variables. These variables were not added as latent constructs due to the few and less cohesive items. Game time consisted of the amount of play (in weekly hours) and years of play. Meanwhile, game experience encompassed belonging and skill, such as gamer identity, perceived expertise, and frequency of competitive play.

3.3.2 Responses to Toxicity – The Role of Toxic Exposure and Gameplay Experience

To answer the first group of research questions and measure the effect of toxic exposure and game experience on coping as victims and bystanders, I conducted structural equation modeling (SEM).

Participants rated how frequently they relied on 12 different kinds of responses to toxicity as bystanders and as victims. Following previous empirical work, I defined four response categories as *productive confronting*, *social withdrawal*, *game withdrawal*, and *escalating responses*. Figure 3.3 shows the self-reported frequencies of each response type by victim and

bystander role.

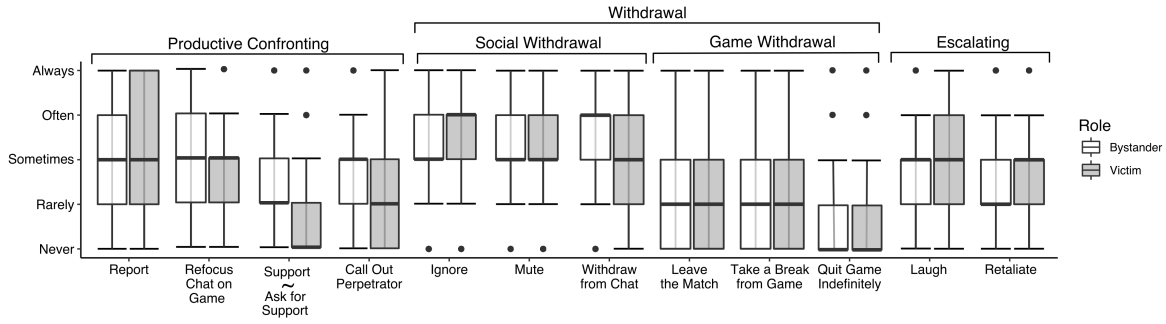

Figure 3.3: Bystander and Victim Responses to Toxicity by Type.

3.3.2.1 Measurement Model

First, I conducted a pooled confirmatory factor analysis (CFA) to validate the measurement model of the structural equation model (SEM). The CFA contained all latent constructs, including indicator (bystander exposure, victim exposure) and outcome (response categories) variables.

The first model distinguished and empirically tested role-specific coping, with four corresponding constructs separately for bystander and victim responses. This, however, produced an inadmissible solution with Heywood cases. Negative variances and latent factor correlation exceeding 1 revealed that some bystander and victim response categories were indistinguishable. This result indicated that responding to toxicity may not be role-based; instead, it is a behavioral strategy. Thus, in connection with RQ1, which investigates the differences in how bystanders and victims respond to toxicity, the data did not show significant differences between bystander and victim responses. To reflect this finding, coping factors were re-specified as fully strategy-based constructs: I kept the previously introduced four response categories, namely *productive confrontation, social withdrawal, game withdrawal, and escalation*, but no longer examined them separately for bystander and victim roles. To reduce redundancy and avoid residual correlations between matching indicators, bystander and victim response pairs were aggregated into composite measurements. Furthermore, I

excluded "ignoring" from social withdrawal and "laughing" from escalation due to low and inconsistent factor loadings.

Aside from the χ^2 test statistic, the indices indicated a good fit: $\chi^2(241, N = 576) = 562.265$, p-value < 0.01, Comparative Fit Index (CFI) = 0.97, Tucker-Lewis Index (TLI) = 0.97, Root Mean Square Error of Approximation (RMSEA) = 0.048, with 0.043 lower and 0.053 upper limit, and Standardised Root Mean Square Residual (SRMR) = 0.070.

For measurement reliability, I calculated internal consistency (omega (ω) coefficient) and composite reliability [21]. Except for productive confronting ($\omega = 0.63$, CR = 0.63), ω and composite reliability of each latent factor exceeded 0.7 ($\omega \geq 0.79$ and CR ≥ 0.83).

Convergent validity was measured using the average variance extracted (AVE), and discriminant validity was tested using the Heterotrait-Monotrait (HTMT) ratios [90]. Factor loadings were strong and significant, with no cross-loadings and all items loading on the intended factor. AVE exceeded the suggested 0.5 threshold for all constructs, except for productive confronting, which was slightly lower (AVE = 0.46). All HTMT ratios were below the strict 0.85 benchmark. The latent constructs, reliability and validity tests, and correlations are summarized in Table 3.2.

3.3.2.2 Model Identification and Evaluation

Extending the measurement model, I added *game time* and *game experience* composite measures and structural paths. I modeled game time as an antecedent of game experience and toxic bystander and victim exposure; and game experience and exposure as predictors of response strategies. Demographic control variables were added for age, gender, and ethnicity. Figure 3.4 [left panel] illustrates the proposed model.

Aside form the $\chi^2(372, N = 514) = 657.543$, p-value < 0.01 statistic, all other fit indices indicated an excellent model fit: Comparative Fit Index (CFI) = 0.97, Tucker-Lewis Index

	Factor	Factor loadings	ω	CR	AVE	1	2	3	4	5	6
1	B Toxic Exposure	≥ 0.72	0.88	0.88	0.60		0.78	0.31	0.06	0.01	0.38
2	V Toxic Exposure	≥ 0.66	0.84	0.87	0.57	0.74		0.42	-0.07	0.09	0.47
3	Productive Confronting	≥ 0.57	0.63	0.63	0.46	0.32	0.42		-0.04	0.40	0.41
4	Social Withdrawal	≥ 0.67	0.80	0.83	0.71	0.05	0.06	0.19		0.43	-0.36
5	Game Withdrawal	≥ 0.66	0.79	0.83	0.63	0.07	0.09	0.37	0.40		-0.16
6	Escalating	≥ 0.88	0.86	0.90	0.82	0.38	0.45	0.21	0.36	0.12	

Table 3.2: Bystander and Victim Exposure and Response Categories to Toxicity. Reliability and validity tests, and correlations. Off-diagonal elements are HTMT ratios in the lower triangle and correlations between constructs in the upper triangle.

(TLI) = 0.98, Root Mean Square Error of Approximation (RMSEA) = 0.039, with 0.034 lower and 0.044 upper limit of the 90% confidence interval, and Standardised Root Mean Square Residual (SRMR) = 0.073. The variance explained by all of the structural predictors was 34% for responding with escalating toxicity, 21% for productive confronting, 12% for game withdrawal, and 11% for social withdrawal.

Reviewing the structural paths, game time predicted game experience ($\beta = 0.39$, $p < 0.001$) and both bystander ($\beta = 0.24$, $p < 0.001$) and victim ($\beta = 0.25$, $p < 0.001$) exposure. Game experience had a negative impact on social withdrawal ($\beta = -0.09$, $p < 0.05$) and game withdrawal ($\beta = -0.17$, $p < 0.001$), and a positive impact on active coping strategies: productive confronting ($\beta = 0.27$, $p < 0.05$) and escalation ($\beta = 0.29$, $p < 0.001$). While bystander exposure only affected social withdrawal ($\beta = 0.35$, $p < 0.001$), victimization appeared to be the strongest psychological trigger predicting all four response strategies: social withdrawal ($\beta = -0.32$, $p = 0.001$) and game withdrawal ($\beta = 0.24$, $p < 0.05$), productive confronting ($\beta = 0.38$, $p = 0.001$) and escalation ($\beta = 0.37$, $p < 0.001$).

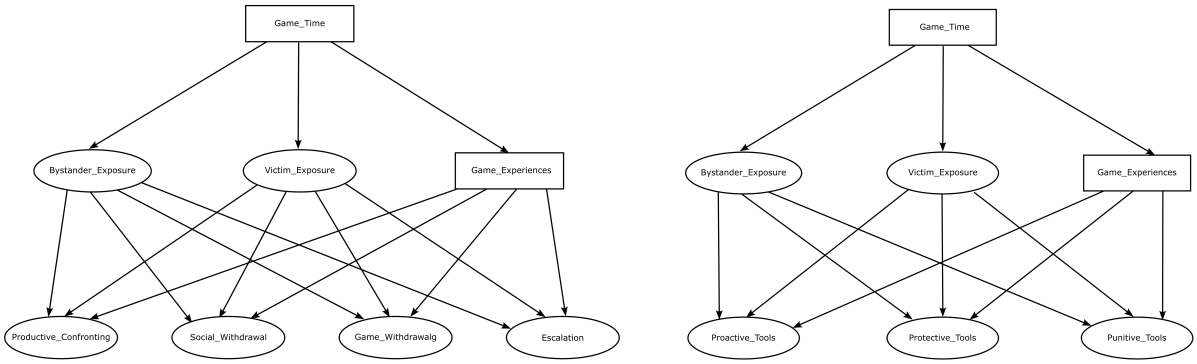

Figure 3.4: Proposed SEM models. The impact of bystander exposure, victim exposure, and game experience on response strategies [left], and on perception of moderation tools [right]

Among the demographic variables, only gender was significant, showing that female and non-binary players withdraw more often than men both socially ($\beta = 0.19$, $p < 0.001$, and $\beta = 0.09$, $p < 0.05$, respectively) and from the game ($\beta = 0.20$, $p < 0.001$, and $\beta = 0.14$, $p < 0.05$, respectively). Further, female players escalate toxicity less frequently than male players do ($\beta = -0.19$, $p < 0.001$). SEM results are represented in Figure 3.5.

3.3.3 Perception of Moderation Tools – The Role of Toxic Exposure and Gameplay Experience

Game companies offer a variety of moderation tools for managing disruptive player behavior in online games. Figure 3.6 shows the perceived effectiveness of the 12 standard moderation tools grouped into three categories: *proactive*, *protective*, and *punitive* tools. On average, participants rated industry tools as somewhat effective (Mdn= 2, IQR= 2). The most valued tools were muting (Mdn= 3, IQR= 1) and blocking (Mdn= 3, IQR= 1), followed by the punitive solutions: banning (Mdn= 3, IQR= 2), reporting (Mdn= 3, IQR= 1), and warning (Mdn= 2, IQR= 1). Proactive tools were seen as less effective, with norm-setting measures (EULA, Code of Conduct, and priming) rated as ineffective (Mdn= 1, IQR= 1 for all three).

To answer the second set of research questions and measure the effect of toxic exposure and game experience on the perceptions of moderation tools, I conducted a second structural

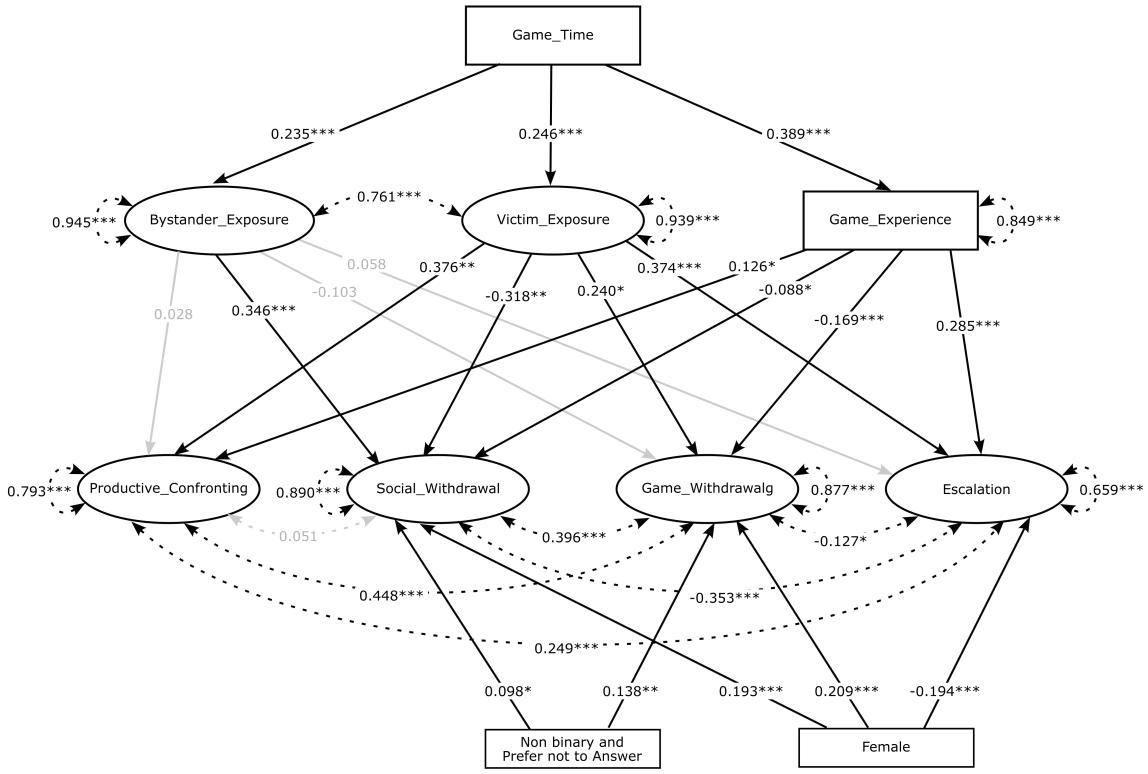

Figure 3.5: SEM results. The impact of bystander exposure, victim exposure, and game experience on response strategies. ($* < 0.05, ** < 0.01, *** < 0.001$)

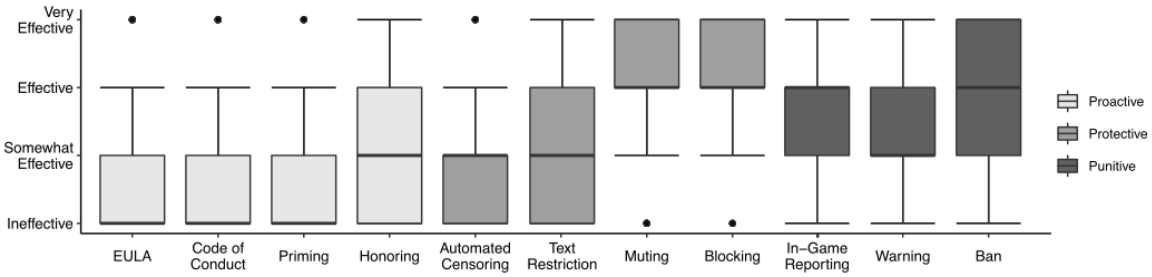

Figure 3.6: Perceived effectiveness of moderation tools.

equation modeling (SEM), similar to the first one.

3.3.3.1 Measurement Model

I again start with a pooled confirmatory factor analysis (CFA) to validate the measurement model of the SEM. The CFA included the following latent variables: *bystander and victim exposure*, *proactive*, *preventive*, and *punitive tools*, allowing constructs to freely correlate. The

χ^2 was again significant, but every other fit index showed an acceptable fit: $\chi^2(252, N = 398) = 627.719$, p-value < 0.01, Comparative Fit Index (CFI) = 0.95, Tucker-Lewis Index (TLI) = 0.96, Root Mean Square Error of Approximation (RMSEA) = 0.061, with 0.055 lower and 0.067 upper limit of 90% confidence interval, and Standardised Root Mean Square Residual (SRMR) = 0.085.

Omega (ω) coefficient and composite reliability (CR) were calculated to assess reliability [21]. Aside from punitive tools ($\omega = 0.68$), ω for each latent construct exceeded the required 0.7 threshold. Composite reliability was strong across latent variables. Average variance extracted (AVE) and Heterotrait-Monotrait (HTMT) ratios assessed convergent and divergent validity, respectively [90]. AVE exceeded the suggested 0.5 threshold for all constructs; only punitive tools were slightly below (0.48). However, factor loadings were substantial and statistically significant, with all items loading only on the intended factor, supporting adequate convergent validity. All HTMT ratios were below the strict 0.85 benchmark. Table 3.3 summarizes the results for the latent constructs, reliability, and validity tests, and correlations.

	Factor	Factor loadings	ω	CR	AVE	1	2	3	4	5
1	B Toxic Exposure	≥ 0.76	0.90	0.91	0.67		0.79	-0.03	-0.03	-0.03
2	V Toxic Exposure	≥ 0.74	0.87	0.89	0.62	0.75		-0.02	-0.05	-0.09
3	Proactive	≥ 0.64	0.89	0.88	0.66	0.06	0.06		0.46	0.70
4	Protective	≥ 0.69	0.90	0.84	0.56	0.08	0.07	0.46		0.66
5	Punitive	≥ 0.57	0.68	0.73	0.48	0.05	0.06	0.67	0.77	

Table 3.3: Bystander and Victim Exposure and Groups of Tool Perception. Reliability and validity tests, and correlations. Off-diagonal elements are HTMT ratios in the lower triangle and correlations between constructs in the upper triangle.

3.3.3.2 Model Identification and Evaluation

Although the CFA showed no correlation between toxic exposure and tool perception constructs, indicating that the proposed predictors do not affect the outcome variables, I tested the structural paths. The model used the same indicator variables as the previous one: bystander and victim exposure latent constructs, and game time and game experience composites. Figure 3.4 [right panel] illustrates the proposed model. Age, gender, and ethnicity demographic control variables were also included.

$\chi^2(372, N = 403) = 688.952$, p-value < 0.01 statistic was again significant, however, all other fit indices suggested an acceptable model fit: Comparative Fit Index (CFI) = 0.96, Tucker-Lewis Index (TLI) = 0.97, Root Mean Square Error of Approximation (RMSEA) = 0.045, with 0.039 lower and 0.051 upper limit of the 90% confidence interval, and Standardised Root Mean Square Residual (SRMR) = 0.085. Aside from the small effect of game experience on endorsing punitive tools ($\beta = 0.12$, $p < 0.05$), I find that structural paths are insignificant, with extremely low explained variance in all tools: proactive (6%), protective (2%), and punitive (8%). This result indicates that the perception of moderation tool effectiveness is not related to toxic exposure. Among demographic variables, only gender was significant, suggesting that female players view both proactive ($\beta = 0.20$, $p < 0.05$) and punitive tools ($\beta = 0.23$, $p = 0.001$) as more effective than male players.

3.3.4 The Relationship between Toxicity Responses and Tool Perception

Finally, to answer the last research question and explore associations between how players respond to toxicity and perceive tools, I conducted a confirmatory factor analysis (CFA), freely estimating how the latent constructs correlate, while accounting for measurement error. The model consisted of the four response variables: *productive confronting*, *social withdrawal*, *game withdrawal*, and *escalation*, as well as the three tool constructs: *proac-*

tive, *protective*, and *punitive tools*. The model showed an acceptable fit despite the poor $\chi^2(191, N = 380) = 591.163(p < 0.01)$ value, with Comparative Fit Index (CFI) = 0.91, Tucker-Lewis Index (TLI) = 0.91, Root Mean Square Error of Approximation (RMSEA) = 0.074, with 0.068 lower and 0.081 upper limit of 90% confidence interval, and Standardised Root Mean Square Residual (SRMR) = 0.090. Omega (ω) coefficient and composite reliability (CR) were calculated to assess reliability [21], and average variance extracted (AVE) and Heterotrait-Monotrait (HTMT) ratios were computed for convergent and discriminant validity, respectively [90].

Diagnostics showed strong reliability and validity for most latent constructs. Values for productive confronting indicated lower reliability ($\omega = 0.54$ and CR = 0.47) and convergent validity (AVE = 0.37). Omega and AVE were slightly lower than desired for punitive tools as well ($\omega = 0.68$, AVE = 0.48). However, considering that this is an early-stage exploratory work and the factor loadings were high and significant, with all items loading on the intended factor, I deemed these latent factors acceptable. The results are summarized in Table 3.4.

The factor analysis revealed meaningful associations between response categories and tool perception. I found *significant associations between proactive confronting and all three groups of moderation tools*: proactive ($\rho = 0.28$, $p < 0.001$), protective ($\rho = 0.14$, $p < 0.05$), and punitive ($\rho = 0.19$, $p < 0.05$). Furthermore, social withdrawal was associated with higher endorsement of protective ($\rho = 0.18$, $p < 0.05$) and punitive tools ($\rho = 0.22$, $p < 0.001$), while game withdrawal positively correlated with viewing proactive tools as more effective ($\rho = 0.33$, $p < 0.001$). Figure 3.7 illustrates the relationship between various response strategies and how players perceive tools.

	Factor	Factor loadings	ω	CR	AVE	1	2	3	4	5	6	7
1	Productive Confronting	≥ 0.54	0.54	0.57	0.37		0.04	0.44	0.37	0.28	0.14	0.19
2	Social Withdrawal	≥ 0.72	0.76	0.77	0.62	0.24		0.43	−0.33	0.10	0.18	0.22
3	Game Withdrawal	≥ 0.72	0.78	0.79	0.64	0.35	0.42		−0.13	0.33	0.06	0.13
4	Escalating	≥ 0.87	0.84	0.88	0.79	0.26	0.32	0.10		−0.06	0.00	0.01
5	Proactive Tools	≥ 0.69	0.69	0.82	0.55	0.11	0.07	0.29	0.05		0.49	0.76
6	Protective Tools	≥ 0.71	0.90	0.83	0.55	0.12	0.14	0.16	0.04	0.44		0.64
7	Punitive Tools	≥ 0.58	0.68	0.73	0.48	0.22	0.22	0.09	0.12	0.65	0.76	

Table 3.4: Response Categories and Groups of Tool Perceptions. Reliability and validity tests, and correlations. Off-diagonal elements are HTMT ratios in the lower triangle and correlations between constructs in the upper triangle.

3.4 Discussion

3.4.1 Key findings

1. Young adults cope with toxicity similarly to other previously studied player populations, preferring avoidance strategies over productive ones, such as confronting the perpetrators. Withdrawing from the game and supporting/support-seeking are the least common strategies.

2. Responding to toxicity is not based on transient bystander and victim roles; instead, it appears to be a higher-level strategy.

3. Game experience and victimization drive active coping strategies, including both escalation and productive confrontation. They also predict withdrawal, however, in the opposite direction. Although more experienced players chose social disengagement and

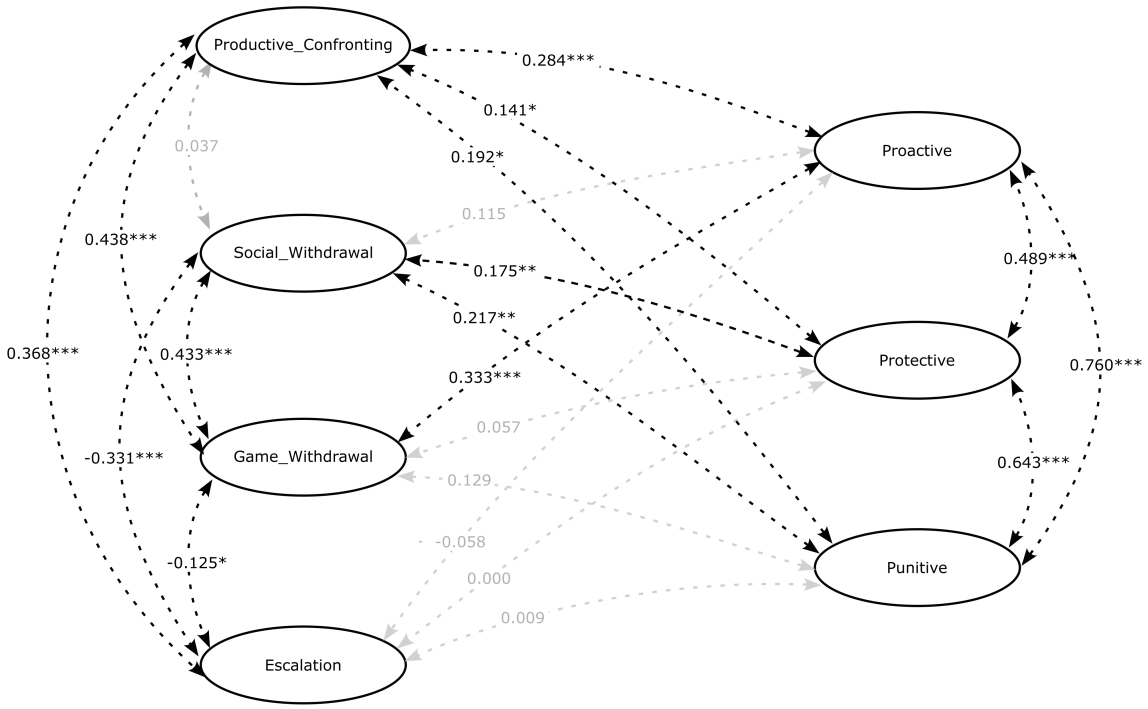

Figure 3.7: Correlations between response categories and tool perception. ($* < 0.05, ** < 0.01, *** < 0.001$)

game withdrawal less frequently, victim and bystander exposure may counter this effect. While bystander exposure increases social withdrawal, victim exposure leads to both social and game disengagement.

4. Interestingly, toxic exposure has no impact on how players perceive the effectiveness of moderation tools. Instead, I found that tool endorsement reflects individual coping styles: (i) players who tend to withdraw from the game view proactive tools as more effective, (ii) those preferring social disengagement value protective and punitive measures, while (iii) players favoring productive confronting endorse each group of moderation tools.

3.4.2 Responses to Toxicity

Young adult players report that toxicity is pervasive in multiplayer online games, with less egregious forms being more frequent, including embarrassment, offensive names, and trolling

or griefing. Their response strategies follow similar tendencies to those of previously studied player groups [2, 12, 72, 212, 235]. The most common responses are low-effort avoidance strategies, such as ignoring or muting the offender and withdrawing from chat. Followed by productive approaches (reporting, refocusing chat to the game, and calling out the perpetrator) and, just as frequently, escalating responses (laughing, responding with toxicity). Similar to earlier results [39, 72, 182, 235], support seeking and withdrawing from the game were the least frequent responses to toxic behavior.

Although previous work suggests that bystanders and victims respond to toxicity differently, while bystanders have the capacity to remain calm, victims often resort to retaliatory aggression [37, 120], I found no evidence of role-based coping. Instead, the data suggests that responding to toxicity is a higher-level behavioral strategy [72], driven by both exposure history and game experience. With more time spent on gaming platforms, players not only develop affinity and skill, but they might also experience toxic behavior more frequently.

Victimization is the strongest psychological trigger, which prompts a broad range of coping strategies, including opposite ones. It activates a fight-or-flight response, spurring players to either openly confront harmful conduct or log off. Victimhood drives both constructive and destructive forms of active coping. However, it also leads players to disengage from the game itself [2, 82, 235]. Why players retreat or fight back can have various reasons: personal disposition (extroversion vs. introversion), the context and severity of harm, and the level of game experience.

Game experience also activates all four kinds of response strategies. While both victim exposure and game experience seem to drive active coping, as opposed to victimization, cumulative experience may lead to less withdrawal. Those who persist build resilience [82, 141, 149] and move from passive disengagement to active forms of coping. With increased experience and skill, players become members of the "in-group". This confidence and stronger sense of belonging may empower players to confront toxicity rather than quietly pull out of chat

or game. Active coping, however, is not always constructive; it can mean either retaliation or productive confronting, such as calling out the perpetrator or supporting the target. Power may be perceived either as an opportunity to take advantage or as a responsibility to help [266].

These patterns suggest that new, inexperienced players are the most at risk. They are frequent targets of embarrassment and skill disparagement [70, 206, 207], and appear to respond with social and game withdrawal when victimized. Those who survive this novice period become more skilled and choose to fight back rather than retreat. Withdrawal, including social disengagement, taking a break, and abandoning the game entirely, suggests that companies have a clear financial interest in keeping players safe. When a growing number of social games compete for a player base, losing players to toxic conduct seems especially regrettable.

Importantly, bystander exposure only impacts social withdrawal, perhaps indicating that indirect experience reduces social comfort. However, while cumulative bystander exposure may erode perceived social safety, it alone doesn't trigger active coping responses or game withdrawal. Through the repeated cycle of witnessing toxicity and responding with social disengagement, it may be the case that witnessing toxicity reinforces bystander inaction.

Taken together, both game experience and victimization predict more frequent retaliation, confirming the fluidity of victim and perpetrator roles [37, 38, 75, 120] and cementing toxic norms by maintaining the cycle of harm [2, 11, 206]. Similarly, bystander exposure increases social disengagement, leaving toxicity unchallenged and victims to cope alone. However, victimization and game experience also seem to build resilience and improve players' skills in productively challenging toxicity. Here, additional longitudinal or case study research is needed to disambiguate the specific in-game contexts in which avid players and victims contribute to or resist toxic behavior.

3.4.3 Perception of Moderation Tools

Young adults' attitudes toward the variety of tools offered by the game industry are decidedly lukewarm. In agreement with previous research [2, 72, 213, 235], among the protective mechanisms, muting and blocking are the two most appreciated tools, providing players agency and immediately removing them from harm's way. The instruments of core moderation – proactive and punitive tools – were rated less favorably, with only banning and, to a lesser degree, reporting ranked as "effective", while the norm-setting tools (EULA, Code of Conduct, and priming) were generally seen as "ineffective". Although game experience has a small positive effect on viewing punitive tools (reporting, warning, and banning) as more effective, I found that the perception of industry tools is not related to exposure history and game experience. Neither victimization nor witnessing toxicity seems to change how players perceive moderation tools. While I found that bystander exposure increases social withdrawal, this outcome doesn't result in viewing the provided protective tools as more or less efficient. Similarly, increased victimhood doesn't impel players to form a more favorable opinion about the tools that effectively protect them or dismiss the ones that fail to do so. And, although more experienced players choose social disengagement less often, they don't view protective tools as less valuable. Instead, regardless of exposure history and experience, young adults seem to value the tools that provide them with control [193] or have a visible impact, like muting, blocking, reporting, and banning. It is worth noting that many tools considered ineffective, such as policies and priming, may very well have positive effects that are harder for players to perceive. Tool perception also appears to relate to players' habitual coping strategies.

3.4.4 The Relationship between Toxicity Responses and Tool Perception

Young adult players who tend to productively confront toxic behavior value each type of moderation intervention, perhaps signaling higher institutional trust and commitment to the shared responsibilities of creating safe gaming spaces. They hold the offered industry solutions in higher regard and take on the responsibility of calling out perpetrators and supporting targeted players. In contrast, players preferring avoidance strategies view the moderation apparatus differently. Those who choose social disengagement value protective and punitive tools, which allow them to shield themselves from harm without actively pushing back on toxic behavior, leaving it to the game companies to deal with perpetrators for the most part. Finally, players who would rather log off than endure or actively fight toxicity endorse proactive tools, such as norm-setting policies and priming. When they see that rules are not respected, they leave the game altogether.

Interestingly, there is no correlation between escalation and the perception of any type of moderation tools, e.g, those who tend to answer toxicity with retaliation don't praise punishment more. It may be that while productive confronting and withdrawal are strategic and dispositional, and therefore reflect players' tool appraisal, escalation is an involuntary, impulsive reaction that happens as a result of emotional dysregulation [255]. Although the importance of immediate and transparent feedback on behavior was identified as an invaluable design principle [117], only a few recent industry attempts have been made to implement real-time warnings [3] and, to my knowledge, no games provide instant incentives for prosocial behavior. Importantly, no current tools offer help with impulse control or self-regulation, highlighting a significant design gap and opportunity.

3.4.5 Demographic differences

Aside from gender, demographic control variables included in the analyses were consistently non-significant. Regarding gender differences, the results confirm my previously discussed findings. Female players' coping strategies and tools assessment fit into the above-described tendencies. In agreement with previous work, I found that women tend to respond with both social and game withdrawal [41, 68, 166, 213] more often than men, and they escalate toxicity less frequently. In agreement with what I found for players who prefer withdrawal strategies, females see proactive and protective tools as more effective than male players. These findings suggest that although game experience (skill and belonging) may help players to switch from passive to active coping, for some marginalized communities, open confrontation remains inaccessible.

3.5 Conclusions and Design Implications

This study set out to examine how young adult players respond to toxicity in multiplayer online games, how they perceive the moderation tools available to them, and whether the two are related. Taken together, my results suggest that players' roles in the toxic encounter do not determine how they cope with toxicity. Instead, responding to toxicity appears to operate as a higher-level behavioral strategy: it is dispositional rather than situational, and not contingent on whether one witnesses or directly experiences a given toxic incident. Which strategy a player prefers, however, is shaped by broader factors: accumulated game experience and cumulative exposure to toxicity both influence the coping patterns players develop over time. The alignment between coping strategies and tool endorsement is perhaps the most practically consequential finding of this work. Players do not assess moderation tools in the abstract; they appear to evaluate them through the lens of their own habitual responses to harm. Those who productively confront toxicity invest trust in the full moderation apparatus. Those who disengage – whether socially or entirely – endorse the tools

that match their preferred mode of self-protection. These individual differences in coping styles highlight the necessity of an adaptive and personalizable moderation toolkit that can effectively support the needs of a diverse player base.

For designers and platform moderators, these findings point toward several practical considerations.

1. Purely punitive moderation architecture – the industry default – addresses the preferences of only a subset of players. Players who productively confront toxicity are, in effect, institutional partners in norm enforcement; supporting and rewarding that behavior through transparent reporting, feedback, and visible follow-through could strengthen their investment in the system.

2. Protective tools – muting and blocking – are already the most valued among players, yet they remain largely static in their implementation; improving their accessibility and granularity would serve the large proportion of players who rely on social disengagement as their primary coping strategy.

3. Perhaps most counterintuitively, proactive tools such as priming and reputation systems may be most valuable for players who are at risk of leaving the game entirely – those who withdraw from matches rather than endure or fight back. Targeting proactive interventions toward this group may improve both their efficacy and their perceived relevance.

4. Importantly, bystander exposure increases social withdrawal and appears to increase the bystander effect. Finding ways to effectively encourage bystanders to defend fellow players and help de-escalate toxic situations could be invaluable in cultivating healthy and supportive game communities.

5. Aside from tools that nudge bystander intervention, real-time measures that support

emotional self-regulation may also help players to choose productive forms of active coping rather than imprudent retaliation.

6. Finally, since tool endorsement reflects coping style rather than exposure history, in-game prompts tied to toxic incidents alone may be unlikely to shift player attitudes toward the moderation apparatus. Instead, design efforts may be better directed at the level of player identity and long-term experience.

3.6 Limitations

This study, like any, has limitations. First, a cross-sectional study can only identify immediate associations between variables that, although a prerequisite to causal claims, are not evidence themselves of causality. While SEM is a powerful tool, when combined with a cross-sectional design, it is most suitable for exploratory modeling and hypothesis generation. The modeled causal inferences leave findings vulnerable to spurious correlations, reversed causality, and measurement biases. The contribution of this work is to suggest patterns and hypotheses that warrant further longitudinal and experimental research. A second limitation arises from the composition of the sample population. The participants were primarily young adults (79.2% between 18-21 years of age) of Asian descent (60.1%) who are currently enrolled in college (63.3%). Previous work [123] found that offline culture has a significant impact on perception and behavior in MOBA games. Although the participants live in the same cultural environment (coastal United States), Asian cultural norms may still influence player behaviors and attitudes related to toxicity. More broadly, young adult players may not generalize to other age demographics. My goal in this investigation was to understand more about young adult "Gen Z" players, specifically given their significant investment in online games at a crucial time developmentally for identity formation, peer relationships, and moral reasoning. This focus, however, may compromise its generalizability.

Chapter 4

From Parental Control to Teen Self-Regulation – A Feature Analysis of Safety, Privacy, and Communication Controls in Popular Multiplayer Online Games

4.1 Introduction

Adolescents' online safety and privacy are a prevailing concern. The most common risks of online social platforms are privacy and information breaches, sexual abuse, child sexual abuse, cyberbullying, and harassment [252]. Both parents and children identify predatory encounters as the single most dangerous threat children may experience online [67], with one in every six children facing online sexual abuse before the age of 18 and the majority of the incidents happening to youth over 13 [64]. Multiplayer games engage millions of adolescent

(Gen Z) players [51], who harness these platforms as valuable third places for socializing and exploring new ideas and worldviews [223]. Despite the well-known benefits, online games are not exempt from harmful conduct either. Harmful conduct, or more commonly toxicity, refers to a wide range of verbal and behavioral transgressions [2, 207, 235] that adversely affect the player experience [82, 125, 130, 137] and cause significant mental, emotional, and even physical harm to the players and others in their lives [62]. Harmful conduct has significant consequences for both individual players [48, 254, 268] and the gaming community [11, 73, 91]. Although predatory conduct on game platforms is a growing issue [47], together with other forms of criminal activities, including doxxing, swatting, sextortion, drug trading [47], and extremist radicalization [200], other forms of harmful behavior, such as name-calling, trolling, grief play, and harassment, are more common threats to players of all ages.

Game companies offer age restrictions, parental control, and a variety of in-game player controls to manage safety and privacy, yet they struggle with cultivating thriving communities and protecting their players. Age verification is often bypassed [89] and only about half of the parents set up parental controls on their children's accounts [67]. Players often adopt strategies that do not rely on technological solutions. Aside from muting or blocking the offender, players most commonly ignore perpetrators or accept toxicity as a normal part of online play [12, 39, 72, 188, 212, 235]. They just as often retaliate with matching aggression or intentionally sabotage the game [37, 132], maintaining the cycle of harm and solidifying toxic norms.

Although available support tools affect players' responses to toxic behavior [72, 193], neither in-game safety and privacy mechanisms that structure social interactions, nor the level of control these tools grant players in preventing and responding to harmful conduct, are well understood. Previous research centers around how players cope with harmful conduct, narrowly touching on their perception of in-game safety measures and moderation tools [2, 235]. Other works focus on specific systems, such as privacy and data protection [15, 139], or

individual titles, such as the reputation system in Overwatch [233], reporting [129] and permanent banning in League of Legends [125], children's online safety risks in Roblox [135] and youth moderation experiences in Minecraft [211, 229]. Available parental controls and safety tools in multiplayer online games, and whether these provide age-appropriate and relevant support for teens, are less explored areas [47].

Here, I aim to contribute to this domain by systematically reviewing the components of the safety, privacy, and communication controls in 12 popular multiplayer online games and the parental controls available in 6 associated publishing platforms. I start with access and account registration, then move on to explore parental controls on the platform level, and finally enter the games and assess the preventive safety and privacy settings, as well as three common reactive intervention tools: muting, blocking, and reporting. In the analysis, I adopt the Teens Online Safety Strategies framework [241, 253], which was originally developed for mobile online safety apps, as a high-level organizing framework and Wang et al. [241]'s three axes of feature variation (granularity, transparency, and to a lesser extent communication) for a more in-depth exploration of how these tools are implemented. Although the number of included games and platforms is limited, and some of the analyzed tools have changed since the data collection (e.g., Roblox's age verification process), I believe that these results remain relevant to the HCI community and the games industry alike.

I found that many platforms support parental controls or linking a teen account to a family, even when teens 13 and older can create "adult" accounts. Parental tools, however, are overwhelmingly restrictive, with only a few features supporting monitoring and active mediation. While screentime and purchase settings were designed to be highly customizable, control over social safety and privacy was less granular and often reduced to the binary on/off options or was entrusted to teens to manage, signaling the values of individual platforms. Interestingly, I found the exact scope of parental controls often ambiguous. Parental restrictions do not always apply to games distributed on the platform, especially those developed by another

company, or in the case of social safety and communication controls. In the absence of parental controls, teens are often left in charge of managing their in-game safety.

In-game safety controls have several shortcomings: (1) they are scattered under different options and fail to present a comprehensive system, (2) in many games, these safety options are frequently unexplained, and (3) they often don't provide sufficient granularity. Similar to parental controls, the components of the in-game safety tools are largely restrictive; they help limit communication and social play by means of safety and privacy settings and muting or blocking, but do not support active coping and self-regulation. Here, I will discuss the implications of these deficiencies and provide design suggestions to remedy them.

The contribution of this work is threefold. First, I conducted a comprehensive and systematic feature analysis of the safety controls on 12 popular AAA multiplayer online games and the parental controls on 6 related publishing platforms. I explored how much control these tools provide to parents and teens in managing their social safety and privacy in games, as well as how transparent and unambiguous these systems are. Second, I mapped the safety features to the TOSS framework to better understand what parenting styles and teen self-regulation strategies are supported and to identify new opportunities and design directions. Third, the in-depth analysis of popular games and platforms revealed broader emergent themes characterizing the current safety, privacy, and communication controls in mainstream games. I identified both common issues and viable practices, and propose design recommendations based on them.

4.2 Background

4.2.1 Youth's Online Safety

Adolescence is a crucial developmental phase, marking the transition from the social status of a child needing adult supervision to that of an adult responsible for their behav-

ior [49]. It is a critical period for forming interests, passionate commitments, and social identities [22, 32, 49]. Here, I consider both the early and late periods of adolescence: tweens (roughly 10 - 14) and teens (around 15 - 20). Adolescence begins with the biological changes of puberty, and its early stage involves considerable physical and cognitive growth and is characterized by increased sensation-seeking, risk-taking, and peer influence [52]. Reasoning capacities, inhibitory control, problem-solving, and reflecting on choices, however, mature slowly and continue developing in late adolescence [49, 97]. This developmental gap between biological and cognitive-emotional maturity creates vulnerability for teens, who navigate complex social situations and manage heightened emotions while the neural maturation of underlying control systems responsible for self-regulation and planning has not yet developed [10, 49]. Therefore, it is crucial to establish suitable social scaffolding that supports adolescents' growing need for autonomy, privacy, and respect and allows them increased freedom in their choices [49]. Adolescents are disproportionately at risk in online spaces of harmful content, contact, and conduct [158] with various types of risks, such as problematic internet use [180], cyberbullying [170, 239], sexual conduct [5, 164], and sexually explicit content [192]. Such harms have significant negative consequences for youth, including depression, distress, and self-harm [162, 210]. A recent survey identified bullying, scamming, pedophilia, and drug trading as the most common criminal conduct in games [47]. Less egregious forms of harmful conduct are more common, including flaming, name-calling, harassment, and hate speech [137, 194], as well as game-specific disruptions, such as grief play, inhibiting team, aiding the opponent, or cheating [124, 137].

Youth online safety is a multifaceted endeavor. It involves policies and educational efforts fostering digital literacies, forms of social mediation practiced by parents, teachers, and peers, technical mediation tools extending social mediation with technologies that help block and filter content and regulate behavior, and industry measures integrated into online platforms and games. In this work, I limit the scope of the background to technical mediation and industry solutions native to the platforms and games.

Although youth face similar threats on social media and in online games, due to the different technological affordances, the strategies offenders adopt may vary. Social media platforms support asynchronous and long-lasting communication primarily through text, image, and video messages and posts. These platforms usually allow circulating and re-sharing content, which can reach a wide audience and remain visible for longer. Video game play, however, typically consists of shorter sessions, with both ephemeral teams and co-players and lasting but more closed social groups (e.g., guilds or clans). Many games allow voice communication, but text and voice messages are usually temporary and often contained within a match or game session. Recent work analyzing sexual harm in video games, esports, and streaming found that esports players and streamers often used their fame and influence to coerce victims and bypass building trust. Children were more often targeted than adults, and the offenders preferred "live" methods, such as voice chat, to reduce the digital evidence they left behind [169].

Aside from the structures and modalities of interaction, technology mediates player safety, privacy, and moderation practices, too. Digital solutions to protect participants from online harm and aim to enforce desired online behavior are widespread today. Such technological approaches are traditionally restrictive and privacy-invasive, often limiting access to digital spaces, filtering content, or tracking usage [55]. These solutions are integrated into online social platforms (e.g., social media apps, websites, or online video games) or embedded in parental control tools, which are often third-party applications. Despite the built-in safety mechanisms and offered third-party tools, parents seem to prefer non-technological solutions, such as rule setting and co-viewing content [157]. While about half of the parents use parental controls on their children's mobile devices, only 35% set them up on game consoles [67]. This may reflect a greater concern for social media usage and perhaps a trust in age ratings to find age-appropriate games [67]. A survey study with 330 children (ages 6 - 17) playing multiplayer online battle arena (MOBA), first-person shooter (FPS), massively multiplayer online games (MMO), and other titles, including Roblox and Minecraft, shared even more

concerning results [47]. 57% of children reported that their parents don't monitor their gaming activities, and over 90% claimed that their parents are unaware of parental control options [47].

There is no consensus regarding the effectiveness of adopting technological mediation tools. Livingstone and Helsper [157] found that neither technical nor preferred active co-use appears to be effective in decreasing online risks [157]. Another study showed that blocking access to websites and tracking online activities were inefficient in protecting adolescents against cyberbullying [170]. However, Ybarra et al. [261] found that monitoring, filtering, and blocking software installed on the family computer reduced youth exposure to unwanted sexual content by 65%.

Although a more controlling parenting style, characterized by direct and preventive strategies (e.g., adjusting privacy settings and using parental control features), may reduce exposure to online threats [251], it comes at the cost of suppressing teens, preventing them from engaging with peers and learning valuable coping strategies [251]. In contrast, active mediation (more reactive, instructive, and monitoring-based strategies) has an empowering effect. It allows a certain level of risk and provides an opportunity to develop coping techniques [251]. Decreasing restrictions and increasing deference and trust over time are necessary to support teens' growing need for privacy and autonomy [27, 178]. An overly restrictive approach may negatively impact teen-parent relations and can hinder teens' psychological development. Therefore, researchers suggest a teen-focused approach and emphasize building resilience and gradual self-regulation in online safety [89, 185, 211, 229, 250, 251, 253].

In addition to parental controls, online social spaces, including video games, implement further technological solutions to mitigate harm. Two common intervention tools are age-based limitation and automated filtering. Video game agencies, such as the Entertainment Software Rating Board (ESRB) and the Pan-European Game Information (PEGI), classify video games based on their content to create age-appropriate recommendations. Neverthe-

less, age-based recommendations are easy to ignore, and age verification can be bypassed [89]. Utilizing automated systems to detect and filter harmful content is another widespread strategy (e.g., [28, 177]). However, censoring is often inaccurate due to inadequate datasets, data overfitting, and the lack of context awareness [8, 196].

When preventive strategies fail, online social platforms moderate harmful content and participation through flagging [43] and blocking mechanisms [101], and commonly discipline users by removing offensive content [102] or suspending the perpetrator's account [244]. Such moderation systems are commonly implemented in online games as well, and are occasionally extended with tools that proactively nudge players toward prosocial behavior. In the following section, I review the related work on industry measures integrated into online platforms and games to moderate behavior and reduce harm.

4.2.2 Technological moderation in multiplayer online games

Moderation in online games follows a largely restrictive and punitive approach. Preventive strategies are scarce, and moderation is commonly an afterthought [125, 246]. Networked games encompass complex governing apparatuses that help moderate play and shape participation in a gaming community. It involves human moderators and increasingly utilizes automated solutions [129]. As with legal procedures, moderation mechanisms, at their core, follow a coordinated sequence of regulation, reporting, adjudication, and punishment [130].

Some instruments of the moderation apparatus allow players to address harmful conduct themselves. The most common tools provided to players to control their experiences are the reporting, muting, and blocking systems.

Reporting or flagging, the entry point of moderation, however, is one of the least trusted and most misused mechanisms. Decision-making – whether an automated process or involves human moderators – happens behind the scenes [125, 257], and feedback on the adjudication

is scarce and usually lacks details. Not knowing how the flags were processed and what outcomes they had erodes trust and leaves players questioning the efficacy and fairness of the process and calling for more openness [125, 129, 137, 257]. layers often choose not to report offenders [2, 11].

If the reported player is convicted, the penalty can range from warnings to content removal, communication restrictions, and the suspension of the player's account [125, 130, 257]. These punitive moderation tools are typical in online games despite their negative effect of drawing the perpetrator's attention to punishments rather than the harm they caused and fully neglecting the victim's needs [257].

More democratic intervention tools, such as muting and blocking, provide agency and immediate results in managing player safety. Muting hides verbal aggression for the duration of the game session, and blocking prevents future communication with the blocked player [2, 213, 235]. Despite being well-liked and widely used, such tools of withdrawal cannot address harm, restore trust, or promote a safe community [187].

4.2.3 The extended TOSS Framework

I use the Teens Online Safety Strategies (TOSS) [253] framework as a high-level organizing framework for analysis. TOSS is a theoretically derived system that combines parental control (originating in parental mediation theory) and teen self-regulation. Both components encompass three strategies. Parental control consists of monitoring, restrictions, and active mediation. Monitoring is a passive parental mediation strategy in which parents scrutinize teens' activities after the fact or co-view the content with them. Restriction refers to setting rules and limiting access or activities, e.g., screen time or spending constraints. Active mediation involves discussion and interaction between parents and teens that help build trust and educate about online safety and privacy. The three components of teen self-regulation are self-monitoring, impulse control, and risk-coping. Self-monitoring is crucial

in self-regulation because it raises awareness of one's own actions and motivations. Impulse control, another significant factor, refers to the capability to inhibit short-term desires in favor of long-term outcomes. Risk-coping happens after enduring stressful situations and includes both controlling negative emotions and addressing the harm. I distinguish active and passive coping strategies. Active coping includes, for example, problem-solving and advice- or support-seeking, while passive techniques correspond to avoiding and withdrawing from the problem, without trying to resolve it. Recently,Wang et al. [241] expanded this framework by three orthogonal axes: granularity, transparency, and feedback/communication support [241] to explore how parental control apps were implemented. Granularity refers to the level of detail and control that a feature provides for the parents, transparency and feedback describe how visible each parental setting is to the children, and communication support shows to what extent a feature encourages parent-child communication. Using these three axes, Wang et al. [241] identifies a proxy of best practice characterized by medium granularity, high transparency, and high communication support. I adopt these axes with slight modifications to better align with the characteristics of online games and platforms (See Section 4.3 for more details).

4.2.4 Research Gap

Aside from recent literature reviews [246, 247], existing works on in-game player safety and governance tools primarily explore specific governance mechanisms and, with few exceptions, focus on individual titles. Examples are a content analysis of Codes of Conduct [81], player perception and usage of reporting in League of Legends [129], the disciplinary effects of permanent banning in League of Legends [125], Overwatch's reputation system as a form of discipline [233], and players' punishment experiences [159, 160] Another group of studies examining players' coping strategies with harmful conduct [2, 39, 72, 214, 235] also provides invaluable insight into players' perceptions and attitudes towards player safety and governance tools.

A significant portion of the extant research focuses on a handful of games, with the overwhelming majority featuring League of Legends. Although the findings may be generalizable between titles, there is no comprehensive overview of how intervention tools are implemented across different games and genres. Although adolescents and young adults are the most active social gamers [51], accounting for 28% of players in the USA [99], youth online safety in games has received less attention (e.g., [47, 135, 211, 229]).

Further, no previous research examined the comprehensive suite of safety, privacy, and communication controls in multiplayer online games, including parental controls, preventive safety, and privacy settings that structure social interaction, and the reactive in-game mechanisms designed to support coping with harmful conduct. To fill this gap and explore the landscape of existing parental and player controls in multiplayer online games, and, more crucially, uncover how accessible and practical they are and how much control they provide to teens and parents, I conduct a systematic feature analysis of these tools in 12 popular online multiplayer games and 6 publishing platforms.

4.3 Methods

4.3.1 Data collection

Here, I analyze popular mainstream multiplayer online games rather than a mix of AAA and indie games for two main reasons. First, the AAA companies behind these titles have extensive resources and decades of expertise in game design and development. Many titles on my list are consistently among the top games since their (or the franchise's) debut several years ago. Second, I aim to include titles that impact the broadest audiences and will persist in the future, ensuring that these findings reflect both mainstream standards and remain relevant for a longer time. Nevertheless, in a few games, safety tools have been modified since the data collection. Further, this study is limited to PC games, although many of

the included titles were published on multiple platforms. I also cut the list of games after reaching 12 titles to keep the scope manageable.

To identify a group of popular multiplayer online games, I gathered a variety of online rankings, including official consumer data and market analytics sites (NewZoo, Statista, Metacritic, Steam Charts), as well as online game blogs and websites covering video games, entertainment, and technology (Imagine Games Network (IGN), Polygon, Gameranx, Geek Girl Authority, GameRant, Geeksforgeeks). I searched for the following terms using Google search engine: best, most played, most popular, and bestselling, combined with online games, video games, and 2024. Single-player games and titles with limited multiplayer aspects, e.g., games with local multiplayer mode or friends-only, were excluded. I compiled the selected list of popular games by tallying the titles and merging the rankings. To ensure a more balanced representation, I assigned genres to the games, and after the first four titles, I skipped games from the same genre. Table 4.1 shows the final index of games, listing their genres, publishers, age ratings, and parental control availability. Most of the games are rated suitable for teens, but even the ones rated as mature (17+), including Call of Duty Modern Warfare 2, Red Dead Online, and Counter-Strike 2, are popular among a younger audience.

Next, I created a data collection protocol to review game features relevant to the research questions. The three areas of focus were (1) parental controls, (2) in-game prevention mechanisms, and (3) in-game reactive tools. The in-game prevention included controls of communication, social safety, privacy, and content filtering, while in-game reactive tools comprised muting, blocking, and reporting systems, the latter aiming to ask for institutional help to increase safety. During data collection, I utilized the personal computers in the Games+ Learning+ Society Center. First, I registered youth accounts for 10 and 14-year-old players and parental accounts when required or when parental controls were available. Then I installed the games on the research lab computers and on my personal laptop. Following the protocol, I systematically located and clicked through the specified parental controls and

Genre	Game	Publisher	ESRB	PEGI	Parental Control
Sandbox	Minecraft	Microsoft Xbox	E10+	7	Yes
Platform	Roblox	Roblox Corporation	T	PGR	Yes
MOBA	League of Legends (LoL)	Riot Games	T	12	No
	DotA 2	Valve Corporation	T	12	Yes
Battle Royale	Apex Legends	Respawn / Electronic Arts	T	16	Yes
	Fortnite	Epic Games	T	12	Yes
Team-based FPS	Overwatch 2	Blizzard Entertainment	T	12	Yes
	Valorant	Riot Games	T	16	No
	Call of Duty: Modern Warfare 2 (CoD: MW 2)	Activision	M 17+	18	Yes
	Counter-Strike 2 (CS 2)	Valve Corporation	M 17+	18	Yes
Open World Action Game	Grand Theft Auto V (GTA V)	Rockstar Games	M 17+	18	No
	Red Dead Online	Rockstar Games	M 17+	18	No

Table 4.1: Analyzed Games, publishers, ESRB, and PEGI recommended age ratings, parental control availability, grouped by genre. E10+ - Everyone above 10, T - Teens, M 17+ - Mature contet, PGR - Parental guidance recommended

safety mechanisms in each platform and game, and collected video recordings. Although the primary sources of this analysis are the platform and game artifacts themselves, the recordings let me analyze the safety settings and tools without frequently logging in and switching between games, platforms, and websites, and removed the impracticality of having sizable games installed simultaneously. Additionally, video recordings preserved the state of these systems, in case any of them got updated during this work. However, throughout the iterative analysis, I kept returning to the games whenever additional information was

needed.

4.3.2 Data analysis

I conducted an iterative feature analysis to describe safety tools, including parental controls and in-game controls for preventing and responding to harm. In the first iteration, I systematically identified every feature related to the categories in the research protocol. Following the method of Wang et al. [241], I conducted a form of cognitive walkthrough. However, while cognitive walkthroughs usually aim to find usability problems and focus on well-specified tasks that the evaluator needs to accomplish, here my goal was to explore and characterize the relevant features. To achieve this, I systematically clicked through each available online safety-related setting and interacted with all of the options. I screen-recorded each systematic exploration, using OBS Studio software (version 30.1.0). The steps of systematic clickthrough were repeated to collect data on parental controls provided by the publishing platforms, in-game player controls, including preventive measures: communication preferences, social safety and privacy options, and content filtering, as well as reactive tools focusing on muting, blocking, and reporting systems. I started by examining parental controls, then continued with in-game preventive tools, and finally, I joined the online play to test the reactive features. Reviewing the recordings, I cataloged the features in a Google Sheet and extended the initial protocol to include categories with newly identified ones. Next, I conducted a second round of closer review, taking screenshots and short textual notes of each feature and organizing them on a Miro board. In this iteration, I coded parental control and safety mechanisms according to the extended TOSS framework [241, 253] to explore both the type of these features and how they were implemented. Although this framework was originally developed to analyze mobile parental control apps for teen online safety, TOSS consists of various parenting strategies and self-regulation approaches that can be applied to other domains related to teen safety and technological mediation [253].

First, I affinity-grouped features into categories of parenting styles and teen strategies [253], then on a more detailed level, I coded every feature using Wang et al. [241]'s three orthogonal axes of granularity, feedback, and communication support. Throughout the analysis, I used these categories more broadly. In the original framework, transparency/feedback related to the level of information a feature provided to kids about the applied parental controls. I extended this category to include how transparent parental tools are toward parents and how informative in-game safety measures are for teens. Originally, communication referred to the level of parent-child communication support. This, however, was only provided in parental control features, but was absent from the in-game safety controls. Granularity, the level of control and customization, was used in the same way in evaluating every feature. The analytical codes were applied at the level of individual components rather than the whole parental or player control system. I further coded the features by purpose and functionality, and extended the TOSS groups with teen self-restriction, as many safety measures belonged to this category. Parental controls were not available at the level of games; instead, they were offered by the publishing platforms. Therefore, in games, teens need to set up safety and privacy features themselves. In the absence of parental controls, here, I applied the teen self-regulation half of the TOSS framework, and again analyzed the features by granularity and feedback for a more in-depth understanding of how these systems were designed. Parent-teen communication support was also not applicable on the game level. For both preventive and reactive in-game features, I assessed ease of access – e.g, how fast and intuitively players can open these settings to modify them or reach the reactive tools to respond to harm. For this, I counted the number of clicks to each feature and the required steps to use them, I evaluated internal and external consistency, and information organization.

4.4 Results

4.4.1 Account Registration

Safety begins with access; therefore, I started the analysis by assessing account registration on the 8 distributing platforms and checking whether any parental controls were available. At the time of data collection, all platforms implemented a low-security age gating relying on players' self-declared date of birth (day, month, and year) as standard practice to stay compliant with the Children's Online Privacy Protection Act (COPPA). Steam made this even easier, using a simple checkbox to confirm that players are indeed older than 13. Half of the platforms allowed children 13 and under to register accounts, including Microsoft Xbox, Roblox, Epic Games, and Electronic Arts (EA). Except for Roblox, platforms required parental consent to set up child accounts, which were also subjected to automatic restrictions. For example, Epic prohibits voice chat, free-form text chat, purchases, custom display names, and playing games not developed by Epic; EA constrains online play on child accounts, and Roblox requires phone verification for voice chat and ID verification to access mature-rated experiences. Aside from Battle.net, every platform considers players as adults if they registered a 13+ account. Battle.net uniquely requires parental approval for creating an account for teens under 18. In addition to age verification, platforms unanimously require players to agree to policies, such as Terms of Service, EULA, and Privacy Policy, and commonly ask for an email address. Otherwise, signing up for any of these accounts is a low-barrier and fast process. As a unique and important safety measure, EA prompts players to set the level of account privacy when signing up, by choosing from 4 options: Everyone, My friends (default), Friends of my Friends, and Private.

4.4.2 Parental Controls

Riot Games (League of Legends, Valorant) and Rockstar Games (GTA V, Red Dead Online) had no parental controls at all at the time of data collection; they only allowed 13+ players to

register and left it to the players to manage their online safety and privacy through platform and in-game tools. Every other platform supported the optional linking of a parental account or joining a family and implemented varying degrees of parental controls. However, joining a family as a minor often means losing all control. Except for Electronic Arts (EA), adolescents cannot freely leave a family once they join, and parental controls can be enforced on their accounts. This seems especially restrictive in the case of Steam, which only allows 13+ players to sign up, who would be considered grown-ups on their own Steam accounts if not joining a family.

Parental controls on the six distribution platforms were assigned into four categories: 1) consumption limits, 2) social safety and privacy, 3) content access and filtering, and 4) activity monitoring.

4.4.2.1 Consumption Limits

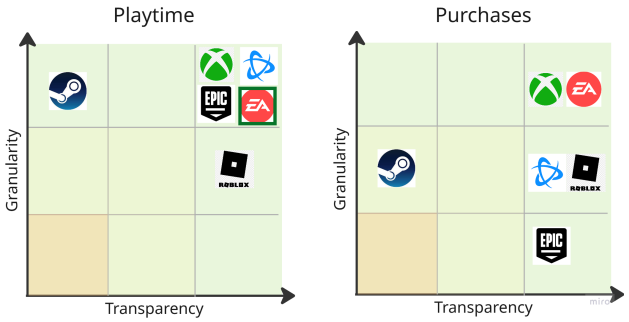

Figure 4.1: Granularity and transparency of consumption controls. Green borders mark controls that teens can adjust.

Two commonly implemented consumption restrictions are screen time and purchase limit settings. Every platform implemented these with high granularity, affording the most nuanced calibration among the parental control tools. For playtime limits, platforms commonly offer a great deal of personalization to match individual needs, ranging from setting no limits at all, through regular daily limits, to scheduling time ranges for each day of the week. Except for Roblox's time management, which offers a simple daily limit only. Uniquely, EA lets teens manage time limits for themselves; although parents can set family playtime in

great detail, for teens (13 -17), it is only a guideline they can exceed. Teens on EA can also set up their own schedules.

Regarding purchase restrictions, platforms commonly allow linking payment options and setting a monthly purchase limit (Xbox, Epic, Roblox, EA) or adding money to the child's balance (Xbox, Battle.net, Steam). Additional spending-related settings are: sending purchase notifications to parents (Xbox, Roblox), and limiting in-game shopping separately from the platform store (Battle.net, EA). Memberships, subscriptions, and gifts are usually distinct from purchase limits and don't require asking for permission either. Both screen time and purchase limitations had medium-to-high transparency and communication. They are typically visible and intuitive for parents, transparently communicating the meaning and effect of each setting. These systems are also transparent to the children. Most platforms display time and spending limits in the child's account and also afford negotiation by sending parents requests to extend playtime or allow purchases. Aside from EA, no platform allows teens to self-regulate their engagement by scheduling playtime. Allowing unsupervised spending was more common; Xbox, for example, lets parents link a credit card to the child's account and turn off purchase approvals. Similarly, parents can set a higher monthly limit on Roblox and set spending notifications to none. Figure 4.1 shows the granularity and transparency of consumption controls.

4.4.2.2 Social Safety and Privacy

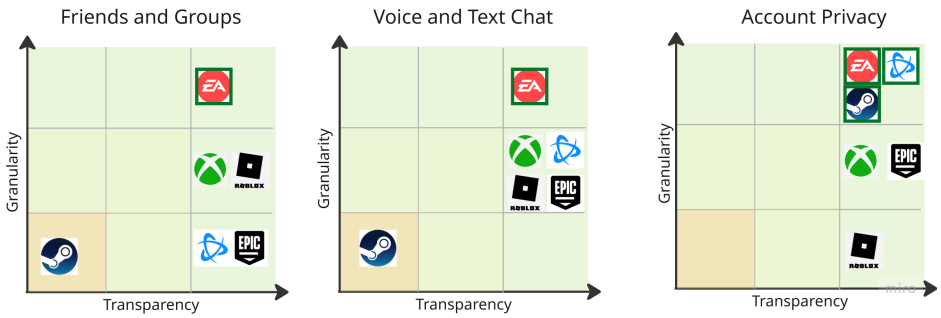

Figure 4.2: Granularity and transparency of social safety controls. Green borders mark controls that teens can adjust.

Among social safety and privacy, I identified several features related to account visibility and structuring social interaction, such as friendship and group management (e.g., sending and receiving friend requests, joining groups, guilds, and clans) or communication settings (voice and text chat permissions).

Account privacy is one of the least supported parental features across platforms. Only the ones targeting child audiences allow parents to control them. Uniquely, Xbox has a variety of related settings, such as limiting who can see the child's profile details, friends and followers, game and app history, or online status. In contrast, Epic has a somewhat hidden selector for account privacy outside of the parental controls interface, and Roblox only allows parents to hide a child's online status and current experience. The rest of the platforms leave account privacy management to the teens.

Compared to consumption limits, social safety settings show less consistency and generally lower granularity among platforms. Overall, there is a shift in moving from parental controls to teen autonomy. In this regard, Steam has the lowest granularity: all social features, including friends, groups, and chat messages, are managed by one toggle, leaving little room for parents to customize their child's social experience. Although it's not clear in the account, social settings on Steam mostly apply on the platform level, and don't affect in-game communication and social privacy. Such settings on Battle.net are more granular, separately enabling or disabling voice chat, text chat, posting on community forums, or joining groups; these, however, are also only platform-level settings. The rest of the platforms let parents tune these features separately, and the settings seem to apply at least to their own games.

In addition to Steam's general restriction on social play, friendship management was included in only three platforms' parental controls. Both Xbox and Epic allow parents to block or accept friend requests. And, while kids can make friends on Roblox without restrictions, parents can see the child's friends list and block individual accounts. Parental controls of managing social groups are also occasional. If present at all, options are limited to allowing

or blocking groups: clubs (Xbox) and friends-only parties (Roblox). Text and voice chat restrictions are similar, but somewhat more granular. A few platforms (Xbox, Epic, and Battle.net) allow parents to select a group of players that children can freely communicate with. The options typically include opening chat to *everyone*, *friends only*, or *no one*. In addition to enabling experience chat, Roblox auto-blocks private messages for accounts of children 13 and under.

The typical lower granularity of parental controls is partially due to giving teens more autonomy in managing their social experiences. In contrast to consumption limits, I found substantial differences in how control was shared between parents and children in managing social safety and privacy, ranging from strictly restrictive solutions that leave parents in control to trusting social safety and privacy settings entirely to teens. Although Xbox's social settings were by far the most granular, these options are only manageable by parents. While the selected rules are highly transparent and kids can review them on their own profiles, they have no opportunity to negotiate or modify these settings. On the other hand, EA allows players to manage friendships, groups, and account privacy without parental supervision. Taking the middle ground, on both Steam and Battle.net, teens can self-manage their social life on the platform once parents turn on the base permission, and they have even more freedom in setting social features in games. In a different, maybe less open, approach, Epic and Roblox allow kids to apply further restrictions. Roblox, for example, permits children to turn group and chat settings off and hide their online and experience status even if parental settings allow these. While Epic players can further restrict their profile visibility. Even when negotiation and self-regulation were not possible, the transparency of social settings towards teens was regularly high. Kids can see these options in their own accounts (Roblox, Xbox), and in some cases (e.g., Battle.net, Epic), parental controls are also accessible in the child's account, providing an opportunity to discuss and co-set rules. Transparency towards the parents, however, was less ideal. The scope of social restrictions set by parents on the platforms was often ambiguous and unexplained. Figure 4.2 shows the granularity and

transparency of social safety controls.

4.4.2.3 Content Access and Filtering

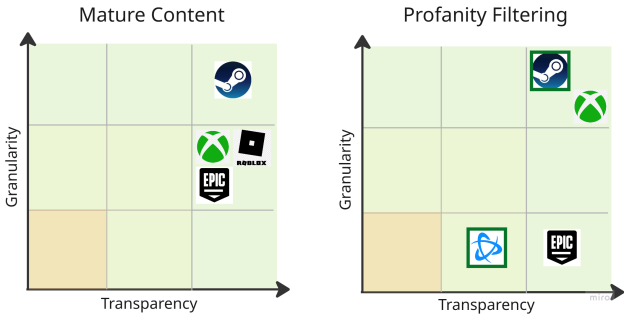

Figure 4.3: Granularity and transparency of mature content and profanity filtering controls. Green borders mark controls that teens can adjust.

Regarding content filtering and mature content settings, parents have varying control. EA and Battle.net provide no means for parents to moderate content access; instead, they fully automate it based on the self-declared age of the account owner. Rather than relying on the child's actual age, Epic, Roblox, and Xbox let parents select a maturity level they find appropriate. Across all the platforms, Steam had the most granular interface, with options to allow or block general mature content, violence, nudity, and adult-only content separately. In connection with transparency and communication, mature content settings are visible and can be further restricted by the kids on both Roblox and Steam. While on Epic, teens could negotiate restrictions by sending requests to their parents to enable games with a higher maturity rating.

Profanity filters are implemented in varying ways, from always-on filtering that neither parents nor children can modify (EA, Roblox) to allowing teens to set their own preferences (Steam, Battle.net). Granularity also varied for both parental restrictions and teen settings. While Xbox offers various filtering options that parents can separately adjust for text chat, text chat with friends, and other media, Epic only provides a toggle to enable or disable censoring (the profanity filter for child accounts cannot be turned off by parents). Teens face similarly diverse granularity, with Battle.net only allowing them to toggle profanity filters,

and Steam supporting broader preferences. On Steam, teens can curate lists of always and never filtered words, opt into unfiltered conversations with friends, and adjust profanity and slur filtering. Figure 4.3 shows the granularity and transparency of mature content and profanity filtering controls.

4.4.2.4 Activity Monitoring

All of the analyzed platforms support sending weekly activity reports to parents, mostly consisting of playtime notices. Xbox sends this document to both parents and children, allowing kids to self-monitor and reflect on their gaming activity. Aside from the email notifications, a few platforms allow both parents and children to track activity on an activity dashboard. For example, on Roblox, both parents and children can see a screen time chart, top played experiences, and friends. Playtime dashboards are also available on EA, both for teens and parents, who can see a family playtime chart with each managed youth account. Players' platform profiles can also be considered as forms of activity dashboards; they usually collect and display information about games played, engagement, and social participation, and can be a great source of self-reflection and monitoring (both parental and teen).

4.4.3 Player Controls of Safety, Privacy, and Communication

Next, I scrutinized the safety, privacy and communication settings in each game, focusing on the level of freedom and control safety features provide for players in customizing their social experiences (granularity) as well as the ease of use and clarity of how they work (transparency). Unlike with parental controls, here the settings are between the player and the platform; therefore, transparency refers to how transparent these features are. I start the analysis with preventive strategies, which aim to protect players from unwanted contact and content, and then move to reactive tools that help respond to harmful conduct.

4.4.3.1 Preventive Strategies – Social Safety, Privacy, and Content Filters

At the level of games, I found that parental controls and platform-level settings do not always apply to individual titles. This is especially frequent regarding social safety and privacy settings. While consumption and mature content restrictions affect the platform and store and can easily prevent excess purchases or block games and experiences with adult content, social features are game-specific. The modalities (voice, text, ping, emote, etc.) and structures (teams, parties, clubs, clans) of interactions are uniquely designed for individual games, making them difficult to manage at the platform level.

Each game in the sample offers various modalities of communication, and commonly implements voice and text chat. In team-based competitive games, where strategy and coordination are crucial, faster forms of context-sensitive alerts and signals are also implemented, including pings (visual signals to teammates displayed on the map to highlight location, objective, or threat) and radio commands (pre-recorded tactical voice messages that help team members coordinate), which were typically available on chat wheels. Aside from strategy-related modalities, emotes and avatar gestures allowing players to express more personalized emotions are also conventional; many, however, are in-game purchases only.

4.4.3.1.1 Voice and text chat. I found several communication-related safety measures, some of which are unmodifiable, hard-coded restrictions, and some that players can customize. Games targeting younger audiences typically restrict access to communication features; e.g., limit voice chat to verified players (Roblox) or have no native voice chat at all (Minecraft). In contrast, competitive games often limit the group of participants to reduce harmful conduct and keep communication strategic and friendly. Except for custom matches and private lobbies, in most of the team-based games, voice chat is limited to the team (Fortnite, Overwatch 2, Apex Legends, Dota 2, CS 2) or to a pre-made party (Valorant, LoL[1]). Whereas, games with mature content, adult audiences, and more realistic open-world set-

[1]The 2026 March update introduced team voice chat in League of Legends.

tings usually allow unrestricted and proximity-based communication with all online players (CoD: MWII, GTA V, Red Dead Online). As opposed to voice chat, text chat is usually less restricted by design, leaving players in control of who they interact with. Aside from Red Dead Online, every game includes text chat, often with multiple channels to message the team, all players, or a single person. Private text messages (whisper) are also commonplace, yet more often limited, due to the risks of harmful contact. For example, Roblox fully restricts private text to accounts 13 and under, while LoL and Fortnite only allow it between friends.

In addition to hard-coded restrictions, games often let players tweak the modalities of communication (text and voice chat) and select the group of people they feel comfortable interacting with: friends, teammates, clans, or, in some cases, opponents too. The granularity of these settings, however, varies considerably. Some games have rather limited voice chat options. For example, LoL players can only talk in premade parties, which are either open for friends and teammates or closed, allowing players to join by invitation only. Having no option to disable voice chat, Dota 2 players can only lower the volume of party or team chat. In contrast, Valorant and Overwatch 2 players can enable different voice channels separately. Other titles (Fortnite, CoD: MWII, GTA V, Red Dead Online) have somewhat more granular voice options and allow selecting the group of people to interact with; options typically include everyone, friends, other groups (team, crew, posse), or no one. CoD: MW II has additional settings related to premade voice channels and proximity chat. Some games (Apex Legends, CS 2) manage voice and text together, making it impossible to disable one while keeping using the other unless muting voice chat by lowering the volume. Text chat settings are often reduced to an on/off toggle (Fortnite, Minecraft, and Roblox), which, in the case of CoD: MW II, closes multiple text channels at once. Overwatch 2 at least allows different text channels to be disabled separately. Only a few titles, including LoL, Dota 2, and GTA V, allow players to select the group of people to text chat with. Some games with no default restrictions on private texting (whispers) allow players to either turn it off

(Overwatch 2) or limit who can send direct messages (CoD: MWII, GTA V).

4.4.3.1.2 Social groups Befriending players and forming groups are also important factors of social safety and privacy, especially because communication can be restricted to friends and party members. Most of the investigated games (Roblox, Overwatch 2, LoL, Apex Legends, DotA 2, CS 2, GTA V, and Red Dead Online) place no restrictions on friend requests; however, in some cases, account-level privacy settings (e.g., Xbox) could affect this. Only a handful of games allow players to set limitations. In Valorant, for example, players can turn on auto-rejecting friend requests, while CoD and Fortnite provide more granular options to manage befriending. Similarly, players can manage group privacy to varying degrees and control who can join their groups and who can send them group invitations. While on Roblox, only friends can join parties, the majority of games (Fortnite, CS 2, Overwatch 2, Apex Legends, LoL, Valorant, Red Dead Online) allow players to open their parties to the combination of *friends only*, *invited players*, or *open to anyone*. Many games also let players block party invites or restrict them to friends, including Valorant, LoL, Overwatch 2, and Dota 2. Among the 12 games, CoD: MW II has the most granular group management with multiple levels of party privacy for both joining and inviting players.

4.4.3.1.3 Privacy settings As discussed above, some other aspects of social privacy, such as account visibility, are commonly managed on the publishing platform, typically by the account holder rather than a parent, even when parental controls are available. A related privacy setting occasionally supports customizing players' display names. For example, options include hiding real names from non-friends (Minecraft), turning on anonymous mode (Apex Legends), hiding display names (Fortnite, Valorant), and selecting what name to show on the avatars' overhead display (LoL, GTA V). Some of the competitive games also allow controlling the visibility of progression and hiding game stats to an extent. For example, occasionally players can turn off sharing gameplay data with third parties (Dota 2) or with other players (Fortnite), and Overwatch 2 allows players to control the visibility of their

career profile and set it to public, groups only, friends only, or private.

In many games (Roblox, GTA V, Red Dead Online), profanity filtering is turned on by default, and players cannot customize it at all. Other titles allow players to turn censoring on and off (Minecraft, Fortnite, LoL), and a few let them further customize it. For example, aside from enabling or disabling the explicit language filter, Valorant players can maintain a muted words list to filter out text messages. Overwatch 2 allows players to choose between mature (light filtering) or friendly (strict filtering) text for each channel (match, team, party, and whisper) separately while providing details on each option, sharing a link to their code of conduct, and reminding players that even with filters on, offensive language is unacceptable.

Additionally to text chat filters, a few games offered clean avatar names. For example, in Fortnite, Valorant, and CS 2, players can switch to neutral tags as display names for those who are not in their party or squad, while LoL allows players to select what to display above players' avatars: player name, champion name, or nothing.

4.4.3.1.4 Access and Transparency Reviewing the in-game controls of social safety and privacy settings, broader trends related to transparency and accessibility emerged. I found that the various components are generally not grouped under a dedicated menu. For example, voice chat-related settings are often grouped with input-output device settings and volume controls in the Audio tab, while text chat, group management, and privacy have less predictable placements. Text settings can be found under various menus, such as Communication (Valorant), Interface (LoL), or Account & Network (CoD: MW II). Both Fortnite and Overwatch 2 organize social settings with more care, and at least group text chat with other social and privacy settings, although voice chat options are still in Audio. Scattered controls are hard to navigate; it takes more mental effort to recall the location of different interaction and privacy options when they are not mapped intuitively. Locating and modifying various safety components on multiple tabs also requires more time. Further,

players may be unaware of some settings because they are hidden among unrelated items or placed in a tab they would not expect to find them. Grouping these settings would help players better understand the available options for adjusting online safety and privacy.

Transparency - the clear and unambiguous meaning of different options - is invaluable for making conscious choices of safety and privacy. Although some of these settings seem trivial, and many are consistent across titles, there still may be ambiguity. Games explain the above-described safety and privacy settings to varying degrees. In some cases, there is no explanation at all (LoL, Overwatch 2 - except for the profanity filter pop-up), while others have inconsistent, occasional tooltips to clarify certain options (Valorant, GTA V, Red Dead Online). In other games, tooltips are more regular and provide consistent help (CS 2, Dota 2), while elsewhere, explanations are always visible under each setting (Minecraft, Roblox). There is, however, an even more beneficial approach. Dividing the screen, some games list the player controls on one side and the corresponding explanation of the selected setting on the other. This structure provides sufficient space for detailed explanations with neatly organized and formatted texts. CoD: Modern Warfare II, Apex Legends, Fortnite, and Red Dead Online (main menu only) all consistently follow this practice.

4.4.3.2 In-Game Reactive Measures - Muting, Blocking, and Reporting Systems

Aside from the preventive safety, privacy, and communication controls, games offer reactive tools that players can use when they face harmful conduct. The three most commonly implemented tools are muting, blocking, and reporting systems. In the following part, I review these systems based on the level of control they provide (granularity) and the transparency regarding their usage and effect. Unlike the more static safety settings that players less often modify, these tools are designed for reacting to harm as it happens. Their usage is more dynamic and frequent, and often happens during a fast-paced game. Therefore, here I also consider how easy and fast it is to access and use these features.

4.4.3.2.1 Muting is one of the most used and favored tools among players, allowing them to swiftly close communication with an offender. Every game that I examined implemented a muting system that players can access at each stage when voice or text chat is available, including lobbies, character selection, and gameplay or matches.

The muting system shows significant external consistency among the games. Except for the two Valve games, Dota 2 and CS 2, all of the team-based titles allow muting teammates before the match, in the lobby, or during character selection. This feature is most visible in the two titles by Riot Games, where the muting option is displayed on the team members' icon (Valorant) or appears when hovering over them (LoL), providing instant access. It only takes one more click in the two Battle Arena games (Apex Legends, Fortnite), where muting is located on a dropdown that opens by selecting the teammate's character. Alternatively, muting is commonly available on the player list by opening the social tab (Overwatch 2, Fortnite), the lobby members (CoD: MW II), or the scoreboard (Overwatch 2).

When entering the match, muting becomes even more consistent across games. Each analyzed shooter (Overwatch 2, Valorant, CS 2, CoD: MW II) and MOBA (LoL, DotA 2) game placed muting on the scoreboard/playerboard. After bringing up the playerboard, players can be muted either directly on the pop-up window listing the players (Overwatch 2, LoL, Valorant, Dota 2) or from a dropdown menu attached to the selected player (CS 2, CoD: MW II). Placing the mute buttons on the scoreboard provides quick access and good visibility.

Less competitive (and often not team-based titles, such as Roblox, Minecraft, Red Dead Online, GTA V) have muting on the online players' list, which can be accessed through different menus: pause menu (Roblox), social tab (Minecraft), online options (Red Dead Online, GTA V). These games also commonly reduce muting to a low-granularity toggle without further customization.

Team-based competitive games, however, frequently give players more control in managing

communication. Here, completely muting someone can hinder play. Therefore, to help reduce friction and noise but keep the communication alive, FPS and MOBA games implemented more nuanced muting. Aside from fully turning off each modality of communication, they allow players to separately mute text, voice, and even pings (Overwatch 2, Valorant, LoL) or to adjust the volume of each player individually instead of completely silencing them (Overwatch 2, Valorant, CS 2). When communication is beyond repair, or the player needs a moment of focused silence, it may take too long to click through players one by one to mute them. To speed this up, many games added a group mute functionality to disable chat with the team or opponents at once. Following consistency, group muting is usually placed on the scoreboard or player list together with individual muting options. Overwatch 2, for example, allows muting match text (all players, including opponents), team text, and team voice separately. In other games, it is simpler and usually one button mutes all channels (LoL, CoD: MW II, CS 2) or enemies only (Valorant).

Additionally, games offer muting by typing commands in the text chat. LoL, for example, allows the use of */fullmute ¡playername¿* to mute players individually, or */ignore all* to mute everyone. These commands, however, are not shown on any in-game interface, and players need to use external resources to discover them.

4.4.3.2.2 Blocking Similar to muting, players rate the blocking mechanism highly, which takes muting a step further and cuts all future interaction with the selected player. It remains in effect until the blocked player gets unblocked. Blocking, however, is not without ambiguity. Players often don't fully understand what it means and whether it prevents meeting blocked players in-game or becoming their teammates in the future.

Compared to muting, blocking a player is usually less urgent, and thus often more hidden on the user interface. Although many games manage blocking consistently with muting, they often don't add this option directly to the scoreboard or players' mini-profiles; instead, it is

typically hidden under an overflow menu to manage screen space and keep the UI concise.

In some cases, blocking is only accessible on the platform level rather than in-game, e.g., in both Rockstar Games titles and CS 2. And although it is rarely explained in the games, blocking usually has a platform-level effect; blocking a player in Overwatch 2, for example, will block their Battle.net account and prevent communication and friend requests in every other Blizzard game as well.

In some games (Roblox, CoD: MW II), blocking is placed together with muting and requires the same steps to access, making it easier for players to remember how to navigate to these options. Aside from the pause menu, blocking on Roblox can be accessed through the overflow menu on each player profile and in some experiences by clicking their avatars.

More commonly, however, blocking is somewhat more hidden than muting. In Minecraft and Fortnite, the initial steps are the same as when muting: by selecting the offensive player on the social tab, but while muting is immediately visible on the player's profile, blocking is placed under an overflow menu. Blocking during a match in Overwatch 2 also takes an extra step, and can be accessed on a dropdown by selecting a player from the playerboard.

LoL and Valorant break from the conventions entirely, and instead of trying to keep blocking together with muting, they don't support blocking during the matches. This approach not only reduces the complexity of the UI but also gives an opportunity for players to reconsider applying this drastic measure. Otherwise, in both games consistently, blocking is available in match finding through the social tab, on the victory screen, or by finding a previous game in match history. Although these steps may take longer, they assume a deliberate decision and don't disrupt the play.

Blocked players can usually be unblocked on their profiles and mini profiles, the same way they have been blocked or via block lists. Block lists are available in about half of the games, and can be accessed in general settings (LoL, Valorant, Roblox, CoD: MW II) or more visibly

on the social tab, together with party members and offline players (Fortnite).

The blocking feature shows less transparency. Only a handful of games inform the players of the consequences of blocking (Roblox, LoL, Fortnite, Valorant) and ask for confirmation (Roblox, LoL). In contrast, in some games (Overwatch 2, CoD: MW II, DotA 2), blocking is a simple toggle, without explanation and confirmation. According to the explanations a few games provide (Roblox, Fortnite, LoL, Valorant), blocking only prevents future communication, connecting as friends, and receiving notifications from blocked players, who are usually not notified of being blocked. Blocking doesn't prevent playing together in the future or being on the same team. An "Avoiding" mechanism, designed to prevent being matched with selected players, was only implemented in Overwatch 2 and Dota 2.

4.4.3.2.3 Reporting Lastly, I analyze the reporting system in the twelve selected games. Reporting is one of the most ubiquitous tools offered by the industry, and the only system players can use to officially ask for support in moderating play and participation. Unsurprisingly, all of the games provided an in-game opportunity to report toxic behavior. Apart from basic similarities, there are significant differences in their implementations. Just like muting, reporting is also available in multiple parts of the game, e.g., before, after, and during matches in competitive titles, inside and outside of experiences in Roblox, and after joining the online game/server in GTA V, Red Dead Online, and Minecraft.

Following an internal consistency, games commonly place reporting options together with the blocking and/or muting features on the player profile/mini profile and in team-based games on the scoreboard. In games that support open-world online exploration (e.g., Minecraft, GTA V, and Red Dead Online), reporting can commonly be accessed by selecting a player from the online players list. Roblox consistently groups reporting with the blocking option and with muting if voice is activated and the experience supports voice chat. Cod: MW II, Overwatch 2, and Fortnite also uniformly place reporting together with blocking, and

keep it available throughout the game. Unlike blocking, reporting is universally available in every phase of LoL and Valorant as well, by directly selecting a player in character selection, and from the playerboard of current and past matches. Only CS 2 and Apex Legends pose some restrictions on reporting. Just like with muting, CS 2 players can only report during a match, and in Apex, reporting is limited to teammates and the player who killed them in the game. In addition to reporting players, Fortnite allows users to report bugs, and Roblox provides an option to report experiences. Furthermore, in Minecraft (Java edition only) and Roblox, when reporting a text message, players can select the offensive lines to submit with the report as context.

Although it is usually fast and easy to navigate to the reporting function, the reporting process itself can be more lengthy. The reporting interface commonly blocks the screen without stopping the game, and the reporting process often requires multiple steps. In the simplest implementation, it is a single pop up with multiple offense categories for players to choose one (DotA 2, CS 2, Lol - in champion select) or multiple reasons (LoL - during and after match,) with the opportunity to include additional details in an open text field (Valorant, CoD: MW II) and mute (Valorant) or block (CoD: MW II) the player together with reporting. Similarly simple, although implemented on a 2 (occasionally 3) step modal dialogue, are the reporting interfaces in GTA V and Red Dead Online. While Roblox and Minecraft (Java ed.) offer a single-page reporting interface, these are more complex, consisting of 2-3 dropdowns with several options, a free text field for explanation, and, in the case of Roblox, a screenshot uploader. The rest of the games guide players through a 2 or 3-step modal dialogue with (Overwatch 2), or without a progress indicator (Fortnite, Apex Legends).

While most of the games keep reporting options and interface consistent, LoL uniquely has three different reporting UIs with options relevant to the current phase of the game. In the champion selection, the 5 options are tailored to toxic behavior typical during this phase, such as refusing to play a position, griefing/hostage taking, or inappropriate name. During

and after the match, these categories get replaced with more relevant ones, including negative attitude, leaving the game, intentional feeding, or cheating. This design offers more helpful and contextual categories, and at the same time reduces the cognitive load on players by minimizing the steps and available options.

Regarding transparency, as seen, the listed categories impact the accessibility of the system, as does being informed of what to expect after reporting someone. Report options are widely different in games; I find examples of short and context-relevant lists (LoL) as well as lengthily options exhausting all kinds of online harm (even when less typical in games) (GTA V, Minecraft), and anything in between. In many cases, the report options are overlapping, lacking categories, or not phrased in an age-appropriate manner. For example, Apex offers options like colluding with enemies and being inactive during a match, both of which can be considered a form of sabotage, yet sabotage is also a distinct category on the list. Although both Roblox and Minecraft primarily target children, the offered reporting reasons are likely phrased for adults. For example, Roblox lists one-word options with no explanation, such as dating/sex, bullying, scamming, or swearing, while the Minecraft Java edition displays categories, such as hate speech, harassment and bullying, self-harm or suicide, drugs or alcohol, child sexual exploitation or abuse, terrorism or violent extremism, non-consensual intimate imagery, and sexually inappropriate content. Aside from harassment and bullying, there are no options for kids to report more typical and less egregious forms of toxic behavior. The Bedrock edition has similar lists, with a few more options to report common cheating and unsporting behavior.

The feedback after submitting a report also varies, ranging from a barely noticeable line on the screen with a report ID (CS 2), through a thank you note with no further explanation on what to expect (Overwatch 2), to implying that there will be feedback if they take actions without mentioning the name of the punished player (Apex Legends). Uniquely, Dota 2 players can view all of their successful reports. And a few games (Red Dead Online, CoD:

MW II) transparently admit that they will not send any notification on the status of the submitted reports.

4.5 Discussion

4.5.1 Summary of Findings

1. Publishing platforms commonly consider 13+ accounts as adult users and apply no restrictions or mandatory parental consent or control. In the sample, Battle.net is the only exception, requiring parental approval for teens (14 to 18). Nevertheless, many platforms allow teens to join a family account or link a parental account, which, except for EA, almost completely removes teen autonomy.

2. Among parental controls, consumption limits were by far the most granular features, providing parents ample opportunities to customize screentime and spending for their children. Other features are less granular and show a steady shift in control. Regarding social safety and privacy, aside from parental supervision, teens gain more autonomy in self-regulation, while content access management shifts from human-involved to AI-supported, leaving fewer options for parents and teens to personalize censoring or simply turn it off. I found that the social safety and privacy-related parental settings don't always apply to the games, and crucially, these boundaries are not always clear.

3. Games protect players by both hard-coded restrictions and by letting them customize their social experiences via adjusting communication and privacy preferences. In addition to managing modalities of communication, friend requests, and groups, games occasionally provide additional privacy features, such as hiding display names and game stats. Games also commonly use profanity filters, which rarely allow players to customize them (Valorant). These options are implemented with varying granularity, from a simple toggle to multiple parameters for better personalization. Importantly, I found that these tools are often scat-

tered across various tabs in the main game options, and many suffer from low transparency regarding how they work or affect the game experience.

4. The design and implementation of reactive tools are consistent among titles, with genre-specific traits. In less fast-paced, open-world titles with no teams, muting, blocking, and reporting are usually grouped and consistently available throughout the game. While in team-based competitive games, these options are designed and organized for easy access and quick use. Besides their regular places, during matches, they commonly show up on the scoreboard, where players can access them with a few clicks. Among the three mechanisms, muting is the most granular, often allowing for muting different channels, groups, and individuals separately. It also enjoys the highest visibility and the fastest access, thereby supporting in-game use with minimal disruption. Blocking is often more hidden, frequently placed in an overflow menu, with limited transparency on how it works. Reporting is also commonly supported in AAA games and is accessible from anywhere in the game. However, the length of the reporting process and the provided report options can significantly hinder usability.

4.5.2 Access and Liability

Accessing online games and platforms has a rather low barrier. Age verification is easy to bypass, and most platforms only require an email address to set up an account. Although most selected titles are free-to-play games with in-game purchase options, price is a potential barrier for kids to access paid games with mature ratings, such as GTA V, Red Dead Online, and CoD: MW 2. The fact that companies only request age verification and that players agree to platform policies shows that access is centered on liability rather than youth safety. By asking players to self-declare their age, companies shift responsibility to teens and parents. Limiting access to players 13 and older is not a decision based on developmental research but a way of complying with the Children's Online Privacy Protection Act (COPPA)

and avoiding its complex safety and data privacy regulations protecting minors. Similarly, platform policies are lengthy texts written in high legalese [23, 238]. They aim to protect companies' interests and intellectual properties rather than to inform players [23].

For adolescents 13 and older, parental controls are almost always optional; without parental supervision and family rules, they are free to create accounts on most platforms and access games with a range of social features. Many platforms support setting up parental controls and managing family safety on a platform level. However, joining a family as a teen effectively leaves one's game accounts at the mercy of the parent. While it may be less problematic for child accounts, with teens who strive for increasing autonomy, it seems unreasonably restrictive and may deter teens from connecting their accounts to a family. Even if joining a family has its benefits – e.g., sharing games – the current design of these family systems renders teenagers powerless. Platforms should tailor parental controls and family safety features to teens' developmental needs. At least they should allow teens to freely join or leave a family, or better yet, co-set rules of access and participation with parents. Without more balance, either teens have to risk surrendering freedom or parents have to give up technological mediation.

4.5.3 From Parental Control to Teen Self-Regulation

Parental controls were implemented on 6 out of 8 examined publishing platforms. Platforms with child audiences are generally stricter and provide more granular parental controls with less autonomy for kids, including teens. The most detailed and consistent parental tools were consumption limits, allowing parents to control and monitor purchases and screen time in great detail. Although screentime on its own is a less valuable metric [105, 262]. Monthly or weekly activity reports are often largely about playtime alone. Instead of focusing on the quantity – the number of hours teens spend in online games – it would be more beneficial to turn the attention to the quality of social experiences. Many online safety risks, from milder

toxic behavior, such as name-calling or embarrassment, to more serious threats, including hate-based harassment and child sexual abuse, relate to social safety and privacy measures rather than screentime. Yet these features were implemented with less detail, consistency, and transparency on both parental platforms and in games. I found the whole spectrum of granularity and control: from highly customizable parental controls that leave no room for teens' autonomy, through systems that intentionally delegate more responsibility to teens, to parental settings that only affect platform social safety and leave it to teens to manage those themselves in games. A similar trend emerges regarding content access and filtering features, with varying levels of granularity, parental control, and teen involvement. With the addition that automated tools managing access to games and censoring content often leave both parents and teens without choice or opportunities for customization. The shift in control, whether deliberate or simply practical, creates an opportunity for teens to self-regulate their social safety and privacy at the level of games. However, in most cases, this shift is not gradual, and parental mediation vanishes when it matters the most. Considering that both parents and children identify predatory encounters as the most significant online threat [67] and that most of the online sexual abuse happens to teens older than 13 years old [64], providing scaffolded and gradually implemented autonomy in communication and privacy could be invaluable. Additionally, not communicating the exact scope and effect of parental controls transparently may create a false sense of safety for parents and vulnerabilities for teens.

Within games, social interaction is governed by hard-coded design decisions, and by safety and privacy settings players can use to customize their social experiences. Despite the various safety controls, however, many games that I investigated have a toxic reputation. Team-based competitive games, especially, are known for intense gameplay, heightened emotions, and frequent toxic outbursts [116]. For teens whose neural system responsible for self-control and planning is still developing, navigating stressful gameplay and complex social situations may be even harder.

Low-granularity social features, such as managing voice and text chat simultaneously or providing a single toggle to turn communication on and off, leave players with a binary choice between all or nothing. Without social interaction, however, what makes these games different from playing against bots? A more advantageous approach was to allow players to select the preferred modalities of communication (voice, text, command wheel) and the group of people (friends, team, everyone) they feel comfortable interacting with. Adequate granularity makes it possible for teens to fine-tune social participation rather than forfeiting it completely.

Crucially, I found that these settings are hard to discover and sometimes not transparent enough. There was not a single game I examined that grouped and presented all online safety and privacy-related features in one dedicated menu item, echoing previous findings on in-game privacy [15]. Instead, these features are dispersed across multiple tabs and menus, often hidden among unrelated settings. The disorganized components also make it harder to understand how these systems work together and create an infrastructure. Organizing them into a single user interface would not only provide players with easy and intuitive access but also present the bits as part of a coherent system.

The transparency and clarity of these features are equally important. Detailed feedback on how these settings work and to what extent they modify sociality is invaluable for players to make informed decisions about their social safety. Yet, many games provide only occasional or no clarification whatsoever. In contrast, a few games supply a detailed explanation of the currently selected feature in a dedicated section, typically the other half of the user interface. As opposed to the always visible short descriptions next to each setting, this method allows for showing the relevant information only in more detail with better formatting and structure, while reducing clutter. Presenting clear information about the safety and privacy settings is invaluable; however, these are still only visible for those who open the settings. Research suggested to improve the design by using privacy pop-ups to remind players of the available

settings and friendly defaults [71].

Reactive tools - muting, blocking, and reporting - were designed seemingly with more attention to player needs and the context of play. Muting is especially suitable for quickly and effectively restricting communication. While some games reduce player options to toggling voice and text chat together, many provide options for more nuanced communication management, e.g., muting channels separately, group muting, or lowering a player's volume rather than muting them entirely. In contrast, blocking is somewhat more hidden, maybe to subtly control overuse. Regarding reporting, I found that although the report feature was easily accessible, the length of the process (the number of steps and the complexity of report options) often makes it difficult to use during fast-paced gameplay. In agreement with previous work, I found that the reporting options were often overlapping, ambiguous, and difficult to apply [130]

Parent-teen communication support is limited to the platforms and to asking for approval of screentime, purchases, or access to games. Other than this, I found no examples of technological mediation encouraging conversation between parents and teens. This issue relates to the current limited educational aspects of these features. Providing safety tips together with prompting communication in the description of parental controls and in-game safety settings could help educate both parents and teens on game-relevant online safety practices and foster discussions between parties.

Essentially, shifting control to teens is paramount, but it needs scaffolding and to be more gradual than the current implementation, which many platforms and games feature. Ideally, parental mediation should not stop at the platform level, and both parents and teens should be informed about the scope and effect of parental controls and in-game safety settings. The educational opportunity that comes with purposefully organizing and explaining individual safety and privacy features is largely underutilized.

4.5.4 Safety Controls in Online Games and the TOSS framework

Finally, to see which type of parental mediation styles and teen self-regulation practices are supported in parental controls and in-game safety measures, I parallel the identified systems with the components of the Teens Online Safety Strategies (TOSS) framework.

4.5.4.1 Parental Control

Organizing parental controls into the categories of Monitoring, Restriction, and Active Mediation, I find that the overwhelming majority of the tools are Restrictive. These include consumption (screentime and purchases), account visibility, content access, profanity filtering, and social safety management. Although a few related notifications, such as reaching the screentime limit or purchase receipts, can be considered a form of monitoring, all of these settings are primarily designed to restrict participation and access.

Two parental controls fall into the Monitoring category: activity reports and activity dashboards. While sending regular reports via email is common practice, implementing dashboards for teens' gaming activity is rather occasional.

Regarding active mediation, I didn't find any parental tools encouraging co-play or conversations about in-game safety. Although requests, sent by kids to extend playtime, approve purchases, or allow playing a game with a higher maturity level, can be seen as forms of active mediation and negotiation between parent and child.

4.5.4.2 Teen Self-Regulation

Here, I include both parental controls that support teen self-regulation and the components of in-game safety tools. Tallying the safety and privacy features teens can use, I again find that most of them are restrictive. Although self-restriction is not included in the TOSS framework, I decided to add it because of the large number of features falling into this category. Among the parental control tools, some allow teens to apply further restrictions, e.g., on mature

content access or account visibility. Regarding in-game tools, similar to parental controls, social safety and privacy features mostly allow restricting access to communication modalities or groups of players to chat with and limiting what others can see, such as hiding account details, display names, or game stats. Although these tools are unquestionably valuable, fully restricting social interactions is not only detrimental to the player experience but may also hinder building resilience and learning how to recognize and manage harmful conduct.

Self-Monitoring is less supported, and on the platform level, it mirrors the parental controls of playtime dashboards and activity reports sent via email. Additionally, many platforms have detailed player profiles, game statistics, achievements, friend lists, and more. These, however, are more suitable for competitive comparison, progress tracking, and demonstrating belonging than mindful self-reflection. Dota 2 uniquely makes game data available for players on their past in-game behavior, showing how frequently they were reported, commended, or had a toxic vs positive game experience, among other information. This can be an excellent source of self-monitoring and reflection.

Tools that strengthen impulse control are practically non-existent. Detailed explanations of the safety and privacy settings may help teens make more informed decisions and prioritize long-term benefits over immediate needs, and blocking offenders may also be considered a tool for de-escalating conflicts; none of these are explicit tools of impulse control. And neither of the twelve analyzed games offers any immediate support for self-regulation during intense situations that would help players to de-escalate harm or resist responding with matching toxicity.

Risk-coping is relatively well-supported in both publishing platforms and games. However, the majority of these mechanisms support avoidance/withdrawal coping rather than more active advice- or support-seeking approaches. Muting and blocking other players are universally implemented in the analyzed games. They allow players to remove themselves from harm's way. While reporting is the only mechanism allowing active coping in the form of

asking for support and assistance.

4.5.5 Conclusions and Design Recommendations

Systematically analyzing parental controls and in-game player controls of safety, privacy, and communication, I identified both issues and promising current practices. First, the scattered components of the safety controls give the impression that social safety and privacy are indeed more of an afterthought [125, 246] than a core design principle. The inadequate communication of what features are available and how players can customize them can further increase uncertainty [15], and low-granularity features leave players to choose between risking harm or turning off social features. Second, parental controls often don't reflect teens' developmental needs: while parental controls are often overly authoritative on the platform level, parental mediation is seldom supported in games (except for the ones designed for children), leaving barely any opportunities to gradually shift control to teens based on their ability to recognize and effectively manage harm. Third, the overwhelming majority of safety and moderation tools are restrictive. They help players limit their participation as a means of staying safe, and when facing toxicity, they encourage players to disengage (by muting and blocking). However, there are no current tools that support emotion regulation, impulse control, or help de-escalate harmful situations.

Despite the overarching issues, I identified several valuable practices listed below without giving a full account of all of these examples as an inspiration for future design work.

1. Asking players to set up account privacy as part of the registration process (EA), which highlights the importance of safety and privacy. It may be beneficial to extend account privacy settings by customizing communication preferences and profanity filtering, which would allow a quick overview of the basic safety and privacy controls.

2. Implementing parental controls to serve as guidelines rather than hard limitations. An

example of this is EA's parental control on time management, which keeps track of playtime and informs both parents and teens when it runs out, yet teens can exceed it and are also allowed to set up their own schedule. Parental controls regarding account privacy and social safety could also be designed in a similar way. Maybe including an option to turn the strict restrictions into suggestions when parents feel their teens are ready for it. This solution could create an opportunity for scaffolded parental mediation and support teens' need for autonomy and privacy.

3. Clearly communicating the scope and boundaries of parental controls to make sure parents and teens know what additional steps they may need to take for in-game safety. Epic Games, for example, notifies parents when they set up a teen account that chat restrictions will not apply to games owned by other developers. This informs parents that they may need to help set up social safety in games acquired through the Epic store, but developed by another company. An even more useful feature would be to remind teens and parents if platform-level settings don't apply to a newly purchased or installed game.

4. Keeping in-game settings related to social safety and privacy grouped in a dedicated user interface. Although this feature was not fully realized in any games, some had at least part of these settings combined. Overwatch 2, for example, has a social tab under the Options with Privacy and Text Chat settings, and only left managing voice channels separately on the Audio tab.

5. Providing detailed and well-formatted descriptions related to each safety and privacy option, such as in CoD: MW II or Fortnite. Both games have a split UI with the list of options on one side and the details related to the selected option on the other. This solution allows for more details, better visual presentation, and even includes links to safety-related resources. e.g., blog posts on the game's website.

6. Making reactive safety measures accessible, granular, and responsive to the current pace of the game. An example of this is how LoL implements muting, reporting, and blocking features. They keep muting quickly available and granular enough to suit different situations, e.g., to group mute the team, fully mute individual players, or mute text, ping, and emotes separately. Blocking, however, is less visible and not accessible during matches, which keeps the UI cleaner and prevents rage-blocking players. Regarding reporting, Riot uniquely designed different reporting UI with options matching the game phase (e.g., in champion selection or during a match). This solution helps players select a reason faster and potentially with more accuracy. Since reporting is usually a longer process, maybe it would be worthwhile to allow players to put a flag on someone during more intense play and ask them if they want to submit a report at the end of a match or in less demanding phases. This would be less disruptive, allow for more detailed and thought-through reports, and potentially lessen rage reporting, too.

7. Providing feedback on reported cases. A long line of work emphasizes the importance of feedback and transparency of reporting (e.g., [129, 137, 257]), yet many games don't notify the reporting player at all. Dota 2's reporting and behavior system uniquely allows players to see their successful reports on their dashboards with details, such as reporting time and reason.

8. Self-monitoring past in-game behavior. Importantly, Dota 2 keeps track of behavior and communication scores to incentivize prosocial behavior. Although accessing it is less straightforward, on their Steam accounts, players can view the raw data used to calculate these scores. This data includes several important details about past games, such as the number of positive matches, reported matches, abandoned matches, and the number of times a player was reported and commended by others. If formatted better and made more accessible, such reports could serve as a valuable way of self-monitoring. A significant issue is that players are not informed about when they have been blocked

or reported, leaving them unaware that others found their behavior offensive. Showing the summary of the actions others take as a response to their in-game conduct could be a valuable opportunity for self-reflection.

4.5.6 Limitations

This study has several limitations. First, to keep the scope manageable, I restricted the analysis to AAA games published on PCs, and only included 12 titles and their publishing platforms. This scope may limit generalizability; for example, it excludes indie games. However, I aimed to include titles that reach a wide audience and industry-leading publishers that may serve as examples for smaller studios. At the same time, the sampling method introduced potential bias toward Western, English-language discourse about popularity, while commercial rankings may privilege certain genres. Future work comparing these findings with the social safety controls on other platforms, e.g., mobile games or consoles, and with a different group of games, e.g., ones developed by indie studios, would be beneficial. Second, these games and platforms are not static artifacts; some safety features were redesigned and updated during this analysis. While Roblox rolled out an entirely new age-verification system and Riot Games published parental controls just before this work was sent in for review, the general findings here remain meaningful and informative. And third, this work is limited to the feature analysis of platform and game-related safety tools. It is purely an assessment of technology and doesn't involve any key stakeholders, including teens, parents, and game designers. Future work should consider engaging these stakeholders and exploring their experiences using parental controls and in-game safety tools as parents and teens, or uncover motivations, values, and constraints driving design and development.

Chapter 5

Real-Time Toxicity Intervention in Multiplayer Online Games. Players' Acceptance and Attitudes toward AI-Supported Moderation

5.1 Introduction

Multiplayer online games exploded in popularity over the past two decades. The percentage of Americans who play online increased from 18% to 88% since the turn of the millennium [60], necessitating the up-scale of moderation systems. To keep up with the millions of concurrent players, gaming platforms increasingly adopt affordable and scalable AI solutions for moderation [79].

Despite the well-known benefits of online gaming (e.g., [77, 86, 87, 174, 265]), harmful conduct, or more commonly toxicity, remains a prevalent issue, extending to a broad range

of verbal and behavioral transgressions [2, 207, 235] that adversely affect player experience [82, 125, 130, 137] and can cause mental, emotional, and even physical harm to the players and others in their lives [62].

To reduce toxic behavior and protect their players, game companies implemented moderation infrastructures. Rather than supporting victims, many of these systems focus on identifying and disciplining perpetrators [126, 129, 257], revealing a restrictive and punitive design approach [193].

These retributive strategies draw attention to the perpetrator and their punishment rather than the harm they caused and the victim's needs [257]. The lack of transparency and agency in adjudication leaves players frustrated and skeptical about moderation systems [129, 215]. Even tools that empower players, such as muting and blocking offenders [2, 235], fail to address harm, restore trust, or foster a healthy community [187]. Utilizing AI in profanity filtering and in adjudicating reported cases of toxic behavior is widespread today. Although automated moderation can reach a verdict in a matter of seconds, it also follows a retroactive and punitive approach [235, 247]. And hiding the decision-making inside a black box further excludes players from adjudication [129].

While the bulk of previous work on AI-supported moderation tests various machine learning algorithms in detecting verbal [40, 165, 172, 224] or behavioral aggression [13, 24], a handful of recent studies explore opportunities for providing real-time support [76, 104, 193, 202, 216]. The present research aims to contribute to this line of work by exploring player attitudes toward real-time AI-supported mediation in online games. I conducted 5 focus groups involving 28 players to investigate whether such interventions are considered valuable. I focused on what forms of AI-supported mediation players view as appropriate and potentially successful in various situations and contexts. The discussions were based on participants' experiences with moderation and the games they commonly played to understand how the envisioned solutions might alter or disrupt the original player experience. Specifically, I seek to answer

the following research questions:

RQ1 What are players' attitudes toward AI-driven moderation interventions in multiplayer online games?

RQ2 What forms of AI-supported real-time intervention do players envision as acceptable and potentially effective?

RQ2a How can AI be leveraged to support players and de-escalate harmful conduct?

Although the focus groups started with a broader scope of general real-time AI interventions, I guided the conversation toward supportive solutions, including helping perpetrators in self-regulation and exploring opportunities to de-escalate toxic conduct as it unfolds.

I found that players are skeptical of AI-driven solutions. They are most commonly concerned with inaccuracy, the lack of contextual understanding, and human judgment, which often results in unfair punishments in currently implemented systems. While discussing potential interventions, participants frequently gravitated toward traditional punitive solutions. Yet, hard-found supportive ideas promoting de-escalation, bystander encouragement, and self-regulation were the most favored ones across all five groups. Consistent with previous work [76, 193], I found that players are more receptive to mediation that provides them agency and choice. AI interventions integrated into the game world were unanimously preferred over external moderation tools that were considered forced and disruptive. In contrast, utilizing familiar and beloved game characters to support and de-escalate heated moments alleviated participants' opposition to impersonal out-of-context moderation layers.

This work provides further insights into players' perceptions of how supportive AI technologies could help mediate behavior in online games. It discusses novel design directions as envisioned by players exploring real-time AI interventions to de-escalate harmful conduct

and help learn healthy coping strategies to address toxicity in multiplayer online games. The relative novelty of real-time supportive AI mediation provides a unique opportunity to co-design cutting-edge technology in partnership with its users and to set standards of transparency and ethics in shaping the future of governing online communities.

5.2 Related Work

In the following section, I first briefly introduce the components of the current moderation systems, highlighting the issues with the punitive and retroactive approach. I then summarize the existing literature on AI-supported moderation in multiplayer online games and the challenges associated with utilizing AI tools.

5.2.1 Moderation in Multiplayer Online Games

Online social platforms practice custodianship by moderating user behavior and content while balancing company values and interests, legal and moral obligations, and technical feasibility [78]. Signed in 1996, Section 230 of the Communications Decency Act shields online platforms in the US from civil liability for user-generated content published on their sites [94]. Due to the rise in hate speech, harassment, and cyber-aggression on online platforms, this policy is receiving increasing scrutiny and skepticism, with many companies pledging to discipline their audiences and moderate content to meet socially accepted standards [125]. The increasing exertion of governmental regulations adds legal obligations to the economic and moral responsibilities of social platforms, including games [221].

Governance is a broader concept that encompasses a series of coordinated mechanisms, including platform policies and community norms that articulate expected behavior, as well as technological systems that enforce these, such as rules embedded in the game code and player safety and moderation tools [83, 125].

Grimmelmann defines moderation as a set of control mechanisms that structure participation in a community to facilitate cooperation and prevent abuse [83]. Although game mechanics and rules significantly impact player behavior, these mechanisms are rarely designed with prosociality and fairness in mind. Instead, they encourage competition, weak sociality [124, 206, 236], and prioritizing individual progress over sportsmanship [17, 45, 125]. Meta-systems, such as matchmaking [134], ranking [33, 131], and progression [61, 125] systems, are also frequent sources of player grievance and toxic behavior. While these mechanisms are embedded in games, moderation, aiming to reduce harmful conduct, is an external layer.

Moderation systems in online games are commonly criticized for excluding players and following a top-down retributive approach [125, 193, 235, 247] that prioritizes catching and punishing perpetrators [126, 129, 257] and fails to provide support for the victims or deal with the harm. Players are usually tasked with the initial flagging of toxic behavior but are excluded from adjudication and are seldom informed about the ensuing decisions [129]. This lack of transparency erodes trust in the fairness of moderation [129, 215]. Muting and blocking allow players to take control in moderating their experiences by cutting communication with the offender [2, 235]. However, rather than resolving it, these tools conceal the problem along with the toxic conduct, and are unable to restore trust or promote community health [187].

5.2.2 Automated and AI-supported Moderation

Online games today host millions of players who interact concurrently on their platforms, making it impossible for humans to keep up with moderation and necessitating the employment of automated systems. Game companies now increasingly rely on automated and AI-supported moderation, balancing the nuanced contextual understanding of human moderators with scalability and efficiency [79, 216]. Automated tools can identify different forms

of toxicity and take action in a matter of seconds. Newer machine learning and natural language processing methods understand context and identify harm with reliable precision [259]. A significant part of existing research focuses on increasing accuracy and training models for a game-specific context; early work used lexicon-based techniques [231], traditional machine learning models [13, 24], and neural networks [40, 165, 172, 224], while large language models have opened new possibilities for prompt-based toxicity detection that require minimal to no training data [113, 186, 205]. Small language models were also proposed as faster yet similarly powerful alternatives [264].

Automated systems in games moderate play in different ways, including blocking and filtering content, and escalating different forms of punishments ranging from warning messages to communication restrictions and account suspension [125, 130, 258]. While the moderation of player-reported toxicity is commonplace, AI can also proactively identify harmful content or conduct by analyzing the gameplay and chat data. Although text chat moderation is widespread in games, companies now increasingly integrate voice moderation with their games [3]. Activision and Rockstar Games, for example, use ToxMod, a third-party voice moderation solution in Call of Duty and GTA Online, respectively [3, 179]. In addition to being relatively inexpensive, fast, and real-time, AI can alleviate some of the burden for human moderators by preventing and reducing their exposure to harm [218].

5.2.3 The Challenges of AI-supported Moderation

Using AI, however, is not completely free of drawbacks. Instead of mitigating the lack of transparency, fairness, privacy, and autonomy, AI moderation appears to exacerbate these pressing issues [80]. Despite the enhanced accuracy and contextual awareness, AI algorithms, especially deep neural networks, hide their inner workings in a black box, making it hard, if not impossible, to see how the decisions were made [54], amplifying transparency and accountability issues. Current AI solutions in moderation follows the same punitive

top-down approach, excluding players from the adjudication process, effectively reducing them to data labelers [125, 258]. Aside from the lack of adequate feedback on reporting, decisions perceived as unfair also breed frustration among players [129, 216]. Even rapidly evolving systems cannot keep up with the ever-changing creative ways players come up with to evade detection [124, 258]. And while many toxic deeds bypass moderation, some innocent content and friendly banter fall victim to over-censorship. Furthermore, inaccurate decisions made by AI moderation are difficult to contest. Recent work on procedural justice argues that, besides transparency in adjudication, explainability and contestability are also paramount [150]. Governing algorithms require normative decisions that often prioritize a small set of stakeholders [151]. Leaving players out of designing the systems and articulating the values based on what their communities should operate raises further ethical questions.

Although AI has the potential to provide real-time support to victims or nudge bystanders to intervene, it is largely utilized as part of a punitive framework [258]. Despite its known issues, retributive moderation is widespread in online games [247]. However, a handful of recent works explore the potential of leveraging AI in supportive ways [76, 193, 216], and identify alternative approaches to moderation, such as restorative justice [257].

5.3 Methods

5.3.1 Data collection

I conducted a focus group study to explore players' attitudes towards real-time AI-supported mediation and explore what forms of interventions players consider helpful in moderating toxicity, (re)directing attention from punitive to supportive solutions.

Focus groups collect data through group interaction on a predefined topic [171]. It requires participants to interact with one another and the facilitator, and this group interaction generates otherwise unattainable data and insights. Focus groups are particularly beneficial

when the research aims to elicit views and understandings, and observe how these develop in a social context [248]. I selected this method because, rather than eliciting existing knowledge and opinions, I aim to construct and negotiate an understanding of real-time AI-supported moderation in a social context through interaction among players.

I adopted a semi-structured study design and utilized predetermined guiding questions and procedures. Standardization is necessary to achieve high-level comparability across groups. First, I asked participants about their game experiences, e.g., *"What multiplayer online games do you usually play?"*, *"If you could, what common forms of toxicity would you eliminate?"*. Then I started the guided discussion about attitudes toward AI moderation and potential real-time interventions. This open-ended exchange evolved from current experiences to potential improvements and new supportive interventions, with no hard-set questions. I relied only on guiding questions to encourage ideation and steer the conversation in the desired directions. These questions included: *"What are your experiences with AI-driven moderation?"*, *"What would you change about [different systems that they mention]?"*, *"What helps you when someone is toxic to you?"*, *" What helps you if you are tilting in a heated game?"*, *"What would [the proposed intervention] look like? How would it work?"*.

5.3.1.1 Procedure

Participants were recruited through flyers placed on the campus of the researcher's home institution and through visiting undergraduate lectures in different programs. Potential participants were invited to fill out a pre-screening survey that collected data on basic demographics and game experience, including preferred genres (multi-select), games played (free-form response), the frequency of experiencing toxicity (5-point Likert scale), and in-game tools used to protect themselves (multi-select). Eligible participants were 18 years or older, active players of multiplayer online games, familiar with current moderation mechanisms, and comfortable communicating in English. After confirming eligibility, participants received a detailed study description of the research motivation, goals, and procedures via

email and were asked to mark their availability in a shared calendar.

Following institutional ethics approval, data collection took place between August 17 and October 30, 2025. During this period, I conducted 5 focus groups with a total of 28 participants, each of whom received a 30 USD Amazon digital gift card for their contribution. The study took place in person at the Games+, Learning+, Society Center at UC Irvine, with 3 participants joining via Zoom in different sessions. During the focus groups, light refreshments were provided. Each focus group took about an hour and a half.

The sessions started with reviewing the research purpose, study protocol, and informed consent. Then I acquired participants' verbal consent and permission to audio-record the discussion for later transcription. Aside from a fixed set of core questions to maintain comparability, the groups were allowed to pursue their own interests. I only interfered to ensure that the conversation did not stray too much from the research topic and to encourage everyone's participation. During the guided conversations, I took notes of the observed interactions, emotions, and group dynamics.

5.3.1.2 Participants

All participants had extensive gaming experience, and many were students in the Game Design and Interactive Media or Informatics program trained in design practices and game development, which ensured ample insights, rooted in personal experiences and interests. Participants were young adults, between 18 and 22, with an average of 19 (3 only disclosed that they were older than 18). 18 participants identified as male, 6 as female, 1 as nonbinary, and 3 did not disclose their gender. The majority of participants preferred team-based competitive genres, such as Multiplayer Online Battle Arena (MOBA) games (21.1%), Shooters (32.9%), and Battle Royale games (15.8%). Among the most played games, they listed Valorant, League of Legends, Marvel Rivals, Roblox, Overwatch 2, Counter Strike 2, and Helldivers 2.5.1

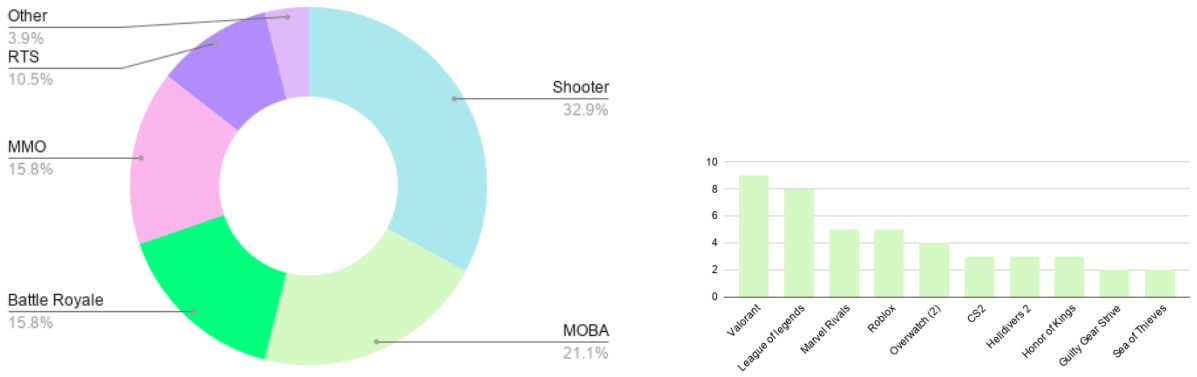

Figure 5.1: Preferred genres [left], Most played games [right]

Participants reported experiencing toxicity in games "often" (M = 3, IR = 1), and used in-game intervention tools to moderate these experiences, including muting (89%), blocking (82%), reporting (82%), and applying privacy settings (61%)5.2.

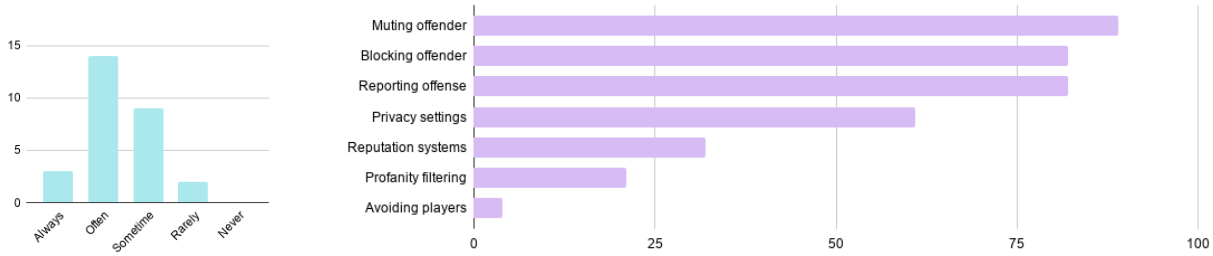

Figure 5.2: Frequency of experiencing toxicity in multiplayer online games [left], Used player controls to prevent or moderate toxicity [right]

5.3.2 Data analysis

A secure online transcription service, Rev.com, was used to produce initial transcriptions, which were manually cleaned for accuracy. Data cleaning also entailed replacing real names and other information that identified participants with pseudonyms. Aside from anonymization, transcription was verbatim, including preserving pauses and hesitation markers (fillers). Audio recordings were destroyed after data transcription and cleaning.

Reflexive Thematic Analysis (RTA) was used for developing themes and patterns of meaning across the data [18]. The analysis was conducted by the sole author; this and the tenets of

RTA made it unnecessary to calculate inter-rater reliability. Reflexive Thematic Analysis (RTA) embraces researcher subjectivity as a resource for analysis. It considers the analytical process inherently subjective and interpretive and posits that coding is neither accurate nor fixed within the data. [19].

I employed an inductive approach and let the codes and themes emerge from the data. Further, I followed a six-phase recursive process introduced by Braun and Clarke (2020) [18]. As a first step, I became familiar with the dataset by reading through the transcripts, producing and organizing high-level notes. In the second iteration, I assigned both semantic codes (e.g., related to different suggestions and design ideas participants discussed) and latent codes (e.g., related to implicit ethical issues). In the following phases, I generated and iteratively refined the themes.

5.3.3 Positionality statement

The results of this analysis are not objective constructs but emerged at the intersection of the data (participants' conversation) and my own subjective experience, domain knowledge, and disciplinary position [19]. I have extensive gaming experience, playing mostly single-player and co-op RPGs and action games, and occasionally multiplayer shooters like Overwatch 2 and Call of Duty. As a woman, I have experienced toxicity and gender-based discrimination in online games, although due to restricting online play to real-life friends only, I was mostly shielded from toxic experiences. I have also intentionally avoided games that have a toxic reputation. I have personal experience with automated moderation tools, and I am skeptical about companies' current privacy, data management, and moderation practices. I am supportive of strong ethical considerations and human oversight in the design and deployment of AI moderation tools, and I am hopeful about AI's potential to realize a supportive framework. My background in HCI and games studies, involvement in the North American academic context, and experience in gaming all shaped the interpretation of the results, but

also granted privilege to Western views and perspectives.

5.4 Findings

As an icebreaker and to ground the conversation in participants' relevant needs, I began each focus group by asking about common forms of toxicity they would like to reduce in the games they played. During the sessions, I kept returning to these issues while discussing potential solutions, their practicalities, and limitations in real games and contexts. Although personal attacks, hate speech, and identity-based discrimination were mentioned, the majority of participants highlighted harm related to competitive play and intra-team conflict, such as bullying newcomers, blaming and insulting each other for perceived performance, and intentionally throwing the match. The discussed moderation ideas relate to these milder forms of toxicity and do not consider more severe forms of harassment, threats, or illegal activities.

5.4.1 Skepticism, distrust, and the lack of transparency

Participants commonly expressed distrust and skepticism regarding AI moderation systems. The most prevalent issues they discussed were inaccuracy and fairness, and, to a lesser degree, privacy and data collection. Transparency was less explicit in the conversations and emerged more as a subtext from how participants talked about AI-supported moderation systems.

5.4.1.1 Fairness and human agency

Participants generally distrusted AI-supported moderation systems, with one exception who acknowledged AI is "*really good at taking in large amounts of data and basically computing it*" (FG4P2), adding that hiring human moderators could be expensive and going through flagged content manually is a tedious job. In contrast to others, he was also more hopeful about AI's capacity to accurately moderate content: "*I feel like AI in games will just be able*

to take all the data from games and just be able to moderate a better level. Not the same, [but] *close to what a human could do*" (FG4P2). The rest of the participants commonly criticized AI's accuracy, its lack of human judgment, empathy, and contextual understanding. For example, not being able to tell apart harassment and friendly banter, or differentiate genuine underperformance and intentional throwing (giving up the match). For example, FG1P7 noted: "*I think it'd be very difficult for AI to distinguish between somebody who's just having a bad game and somebody who's purposely throwing*", and FG3P3 remarked: "[It] *is almost impossible, at least I can't wrap my head around a way, where the AI could truly have the full context of people's relationships with each other, how comfortable they are with each other in saying things to each other*". They mentioned several cases where such limitations led to unfair decisions and concluded that AI "*isn't smart enough to make human judgment*" (FG1P8). Although it can be used to detect " *bad words or uses of cheat programs*" (FG1P8), it is largely unreliable in moderating text chat. They explained how easily players can phrase or spell their insults to remain undetectable, and they cited several cases of overcensorship and wrongful bans "*to the point where it starts debilitating the social game*"(FG5P2). As FG3P6 noted, "*in whisper* [private message] *chat also I was talking to an online friend of mine, and then it was just black. I think it is pretty well known by now that Roblox has started to implement AI. That's a big part of its moderation. So I just wanted to bring that up as a personal demonstration that AI can really just flag innocent stuff a lot of the time, and that's why I'm pretty much against the use of AI excessively in moderation*". Each group highlighted the importance of involving humans in the moderation process and agreed that AI should be mostly used to flag inappropriate content, "*but instead of the AI acting on it, it could send it to a human reviewer*" (FG2P6) and either the moderators or the "*community*" (FG2P2) should make the final decision. Although they rejected the idea of a fully automated process, they recognized the scalability issue of human moderation. As FG1P5 remarked, "*I don't think there's even enough Riot employees in general to be able to review every game that's going on*". Yet, they felt that companies are going in the direction

of leaving humans out of the moderation process, whether they like it or not: "*but I feel like knowing companies, they're just going to make AI automate the entire process, which is what I don't want*" (FG2P2).

5.4.1.2 Privacy and data protection

Participants were also concerned about privacy and data protection practices. However, these issues did not come up as often as accuracy and fairness, and participants' opinions were less aligned. Some were wary of AI monitoring them and collecting their chat or gameplay data, and were concerned that "*some people might see that as an invasion of privacy*" (FG4P2). Others were worried about how real-time chat filtering would affect the game experiences, warning that "*having AI constantly monitor your chat and possibly prevent communications from sent out, especially in the competitive scene, is something I feel could mess up a bit of the game*" (FG4P3). Some were anxious about not knowing the boundaries of surveillance and whether monitoring would still be "*on even when you don't have the app anymore*"(FG4P3). Related to privacy, players were also concerned about data protection practices, and mentioned different forms of misuse, such as targeted advertisements ("*you say* [in game chat] *you're trying to cook later on and then later on just get a random ad about a cooking book*"(FG5P5)) and selling data to third parties (FG5P3). Others were distrustful of companies keeping their conversations for a longer time, as FG5P2 said "*There are voice logs for many games that handle voice communication because you can report it and then they explicitly say like, oh, we'll review the logs, so that implies that they have a record of whatever you say, which is kind of scary*". The ethical issue of players not being informed about how companies use their data was also raised, including the concern that "*they* [the players] *could be the people that AI is being trained by*" (FG2P5). Many of these problems are closely related to transparency and the issue that players are not well-informed.

Some participants were more cognizant of privacy and data collection; not expecting privacy in any online space, and moderating their own behavior and information sharing accordingly.

For example, FG4P2 noted: "*I say, there's no privacy at all. So I am okay with it. I accepted it already* [...] *And every little thing I do, it's being saved somewhere. So I know that's why I'm more reserved when I do stuff just because I know there's no privacy anymore*". And others were less concerned about their data being used to train moderation systems, as long as it was for the "greater good". Like FG5P3 said: "*So I don't mind the data collection regarding the voice logs, and because it's similar to them collecting even the chat logs.* [As long as] *it's not being sold to a third-party license. I don't mind it being used to train their own AI*".

5.4.1.3 Transparency

Transparency was less often mentioned, but as discussed regarding accuracy and privacy, participants showed a lot of uncertainty about how current moderation systems worked and how companies handled data collection and privacy. When talking about moderation in the games they play, they often spoke cautiously and with hesitation, revealing how opaque these tools were regarding whether AI was used at all and how it worked. For example, FG2P2 admitted "*well I'm not sure if Riot uses AI for their moderation, but I've seen cases where they've been flagged for something that was clearly innocent*". While in other cases this unawareness was more subtle, such as " *I think it's just kind* [of] *scary to think about a feature where it's like AI is so advanced, where I guess, maybe be in real-time or something, they can detect those* [chat abuse], *or is it just a bit far stretch*?" (FG5P4), revealing that companies might not communicate transparently that they have already implemented AI-supported real-time text or voice chat moderation/monitoring.

5.4.2 Debating punitive solutions and focusing on the other

Participants' general aversion to AI moderation focused on the limitations and technical feasibility of potential solutions, which prevented them from freely generating ideas. To remove this obstacle and encourage creative brainstorming, I asked each group to imagine a world

without technological barriers. Even so, participants routinely started with punitive ideas that they were already accustomed to and kept returning to those despite being prompted to explore supportive solutions. Punitive moderation ideas were based on existing solutions and included (1) standard moderation (real-time filtering, muting, and banning players), (2) AI substitute (replacing toxic messages and players), and (3) player segregation (rehabilitation and separating toxic players).

5.4.2.1 Standard moderation

Standard moderation was the closest category to the current systems in place. With the perfect AI in mind, participants suggested filtering or muting toxic players in real-time, or blocking communication for every team member until the match ends or for an increasing amount of time. As an example, FG2P2 recommended filtering out toxic lines: "*I feel like in the doom scenario, maybe if the AI detects like, oh, one team is losing and someone's getting heated, you can start having the AI filter messages to stop the actual toxic messages but keep the actual useful communication in there*". While FG5P2 proposed muting the toxic player who " *aren't allowed to engage with voice or chat communication because the game decides like you are not going to be an asset for this team*". In connection to this idea, when things escalate, and a single perpetrator cannot be identified, AI can turn off communication entirely to "*let the whole situation calm down*" (FG3P2). Or as FG5P3 proposed, "*AI can* [...] *just put a chat ban for few minutes and then lift it up and then if it continues, then increase the duration of the ban and then keep incrementing it if it still keeps continuing*". Others pushed back on this suggestion, saying that a short timeout will not change anything, and when it "*runs out, they're going to be even more malicious*" (FG5P5). In addition to communication restrictions, players urged the use of AI to detect when someone is throwing a game. They criticized current systems' inability to recognize AFK (Away from Keyboard) players who pretend to participate by moving their character around to evade detection, but essentially are throwing the game. A participant suggested using match data to detect this

issue: "*So if an AI can look at match history and see, oh this person's doing the same thing where they're just AFK but they're trying to dodge it by just moving around still typing and all that stuff, then if an AI can recognize that and then Chat ban 'em or just ban 'em at all*" (FG3P2).

Increasing and decreasing restrictions based on behavior were another common suggestion. Participants recommended gradually increasing punishment within a single game to correct behavior and determining a penalty based on the player's behavioral history. As an example, FG4P2 recommended: " *I think you could start with maybe just not allowing you to communicate, shutting off text, shutting off voice chat, starting with that and then if they continue, if in response that they start playing bad all of a sudden, they intentionally start throwing the game or they shoot a teammate, then you could escalate*". While FG3P6 suggested using AI "*to assess the rate at which someone's muted per say a hundred matches or something and then decrease or increase the severity of the punishment based off of that every time they act up*".

A more supportive yet still punitive idea was gradually decreasing punishments. Participants suggested sending perpetrators to "*rehabilitation*" (FG1P8) or having them earn back their reputation (FG3P4) before their gaming privileges were restored. "*First of, I think that player has to be at least get punished. So ban for 24* [hours] *of not playing that game, and after that punishment, I think AI could help do something like a rehab session.* [...], *instead of playing the actual game, you have to play 60 hours of that rehab session*" (FG1P8). However, this idea was quickly dismissed, with participants arguing that a troll is not "*the kind of person who would be really receptive to that* [in-game therapy]" (FG1P2). Others suggested incrementally restoring privileges based on good behavior: "*when you come back you have a limited access* [...] *you can't voice chat, you got to play other big games and you might be able to text every once in a while and once you get a few hours, maybe a day of good games, good messages, you can get your chat back or something*" (FG3P3).

While these already common restrictive interventions were among the first things participants explored, they showed no enthusiasm for any of them, and doubted they could solve the core issue of toxicity. They highlighted that those who want to troll others will take the punishment and continue being toxic, and "*even with a perfect AI, you're never going to get someone to fully change their ways, and they're going to realistically be toxic again*" (FG1P5). As FG5P2 noted, bans only limit player engagement, but not "*necessarily improve anything.* [... And] *doesn't encourage me to approach this in a different way*".

5.4.2.2 AI Substitute

AI substitution was another option participants considered and then questioned as a viable solution. One group briefly delved into the idea of hiding offensive chat texts, and "*and then either nobody responds to it, or the AI pretends to be the player it's intended for and then replies back*" (FG1P7). This, however, was quickly rejected. Participants considered it "*an ethical issue if the AI is trying to impersonate a real person*" (FG1P2).

Replacing a toxic player with an NPC was a more popular idea that came up in two groups. Swapping a player with an AI appeared as a win-win solution, because for the team "*at some point an AI player might be more valuable than a griefing player*" (FG1P7), who is "*upset about something, and clearly don't want to play*" (FG1P4). Logging off in the middle of a match and leaving the team in disadvantage is usually penalized in competitive games. Therefore, replacing a throwing or AFK player with an NPC was seen as a resolution favorable to all parties: the team won't be handicapped, the griefer can get out, without being punished (FG1P7), and the opponents are not in disadvantage because of the AI as "*it's not as difficult to adjust the AI's level to the same as the player who's being replaced*" (FG2P4). They further justified this idea, recalling that it was already implemented in Dead by Daylight, where "*you can just leave and then at least the team has a bot that takes a place like a good player*"(FG1P4). The idea of making this replacement optional, either based on a team vote (FG1P7), or on the griefer's decision (FG1P4, FG4P2), was more opposed.

Participant argued that in competitive games "*it feels kind of like a cop-out*"(FG1P7). An alternative version with temporary substitution was also explored. In this proposition, the toxic player would be benched (FG2P2), similar to real-life sports, while an AI NPC would take their spot. By the end of the ideation, participants became wary of being replaced by bots or even playing with them instead of humans. An AI replacement would reduce the social aspect of play, as FG1P2 said, "*there's a certain humanness to playing with humans* [...] *you can learn to read their movements and behavior patterns and everything, and a bot won't really have that*", while the idea of an AI that can mimic human behavior was considered unsettling (FG2P7).

5.4.2.3 Player segregation

Segregating players was another commonly discussed option to moderate toxic players. Participants imagined separating repeat offenders to play either against bots or each other in dedicated toxic servers and lobbies. Based on the example of League of Legends, which punishes toxic players by banning them from ranked games until they win some matches in normal mode, participants in the first group recommended sending perpetrators to bot matches instead. They argued that matching against AI is less harmful than grouping offenders with human players in unranked matches, and that, unlike in League, they should be pardoned based on how they talk and behave rather than on winning (FG1P5). They found this a fitting punishment for trolls specifically, whose "*purpose of being toxic is to get a reaction*" (FG1P2) and proposed keeping them in bot matches until they change their attitude (FG1P3). An alternative solution participants discussed was using AI to track toxic players and match them against each other. This ideation was also based on existing systems; participants mentioned CS:GO (FG2P4) and GTA V (FG2P6) as examples of games that segregate players by behavioral ranks, and a Dark Souls game that isolates cheaters on their own server (FG4P2). They considered adopting such systems in competitive shooter and MOBA games, with AI tracking players' behavior and flagging accounts for toxicity to

group players "*who have a similar mindset of harassing people online together*" (FG1P7). Although they acknowledged the opportunity to "*reward good behavior by putting you with other people who are also playing nicely*" (FG2P2), they more commonly raised ethical concerns and recognized how harmful it might be to force players to join toxic servers: "*I don't know if this is good socially because it's kind of creating echo chambers*", said FG1P7, while FG1P2 noted: "*I can't, don't really imagine it would be fun for an entire server to be bullying each other*".

5.4.3 Discovering supportive solutions and focusing on the self

Participants found it difficult to imagine supportive tools without having seen existing examples. As FG2P7 said: "*I'm having a hard time imagining non-punitive actions that can be taken against people who make a gaming experience worse. I guess I just don't have a frame of reference for it*". Yet, eventually, every group moved from the punitive approach to exploring supportive interventions. They brainstormed several opportunities to support players, and as the discussion evolved, they gravitated towards ideas that were integrated with the games rather than existing as external tools. In the second half of the creative ideation, they revisited previous supportive ideas and reimagined them as integral parts of the games.

5.4.3.1 From incentives to informed self-regulation

Participants envisioned several ways AI could help manage toxicity based on behavioral and performance data, including providing reputation scores, giving strategic tips during play, and managing stress, performance anxiety, and fatigue.

5.4.3.1.1 Reputation and player profile Based on existing reputation systems that allow players to commend one another for sportsmanship, participants envisioned an AI-driven reward system. They mentioned games – such as League of Legends or Dark and

Darker – with player-driven behavior and reputation mechanics. Instead of an optional commendation by players, they proposed a consistent AI-assisted system that tracks and rewards good behavior. They proposed rewarding both successful self-regulation: "*Hey, kudos for you for making your behavior better. Half of a lootbox*" (FG2P2), and helping other players: " *If you help someone out, you're being nice, the game will be like, oh, it'll compliment you, and then it'll give you an experience boost*" (FG4P2). Recognizing that a system built on incentives may not result in a healthier community but "*people just being fake nice*" (FG1P2), participants shifted to reputation systems without rewards. Such mechanisms were imagined as a rating from zero to a hundred (FG4P2) or a healthbar-like slider (FG2P4) publicly showing the level of toxicity. This idea was based on Rainbow Six Siege, as FG4P2 explains: "*They have basically a meter bit that resets every month, and it basically shows how toxic you are, but it doesn't negatively impact your game at all*". Participants argued that such a system would be valuable in other competitive games as well. Unlike reward systems, it would not create unnecessary competition or penalize players with a low reputation. They argued that such an indicator would help them know what to expect from other players and prepare mentally, "*because if this guy is known for being toxic, then I won't take anything he says to heart*" (FG4P5) It was also seen as a potential tool for self-reflection. For some, keeping it in the non-toxic range alone could be a source of self-esteem (FG4P2). Others viewed it as an opportunity to see if their behavior and performance were related, as FG4P4 explained: "*Every once in a while, if I'm just chilling in the game and I'm just curious like, oh, what's my relationship between what toxic I am versus my performance?*".

5.4.3.1.2 Just-in-time guidance Participants argued that tilting and spiraling into a negative mindset due to underperformance were often self-inflicted and preceded being flamed by teammates. As FG4P6 explained: "*a lot of players get the toxicity from their sense of the failure in the gameplay is they feel like, oh, I can lose the game. I didn't do good enough. So I'm just going to lose faith in myself and then try to be toxic*". Tilt can also

occur due to skill disparity between players. To support those who perform poorly or are less skilled, participants suggested providing in-game AI assistance. As FG1P7 explained: "*for the player who's doing really bad,* [the game] *becomes a lot more guided so that they start making better decisions*". They pictured this support mechanism to be similar to tutorials, arriving in the form of small hints and suggestions, but without telling how to play the game or secure a win (FG1P4). Regardless, the ethical question arose as to whether this idea counts as cheating or not, with participants arguing for both sides. Some felt that if help comes only when the player is already tilting, it would just balance out the disadvantage (FG1P7). They also mentioned that hiring real coaches or using third-party tools is common in some games: "*third-party websites or third-party downloads that help you in little ways for League to be able to track certain things. It gives you timers, it helps with accessible information, but to newer players it might not be as accessible to 'em or they might not know that they can access it*" (FG1P5). Others felt that getting any extra help during a match is unfair (FG4P3, FG4P4) and abusable, stating "*I would imagine if the losing team or the toxic team got a boost that would just be abused*" (FG2P2).

As an alternative to in-game tips, participants suggested that AI provide a supportive performance analysis: "*basically identify gently what things they could improve on and then the things they've already done well on*" (FG4P4). In addition to strategic tips, participants also proposed using AI to encourage players, such as " *pop some nice message on the screen to tell the players that you're doing great, your teammates are doing great, let's do this*" (FG2P4). As FG4P2 said: "*I think a compliment goes a long way*".

5.4.3.1.3 Performance and behavior-based nudges Another suggestion was to utilize game statistics to help players overcome performance anxiety and tilting. Many participants recalled games in which a teammate reprimanded them for playing poorly, although they felt they were contributing more than the one pointing fingers. As FG2P2 explained: "*When you're throwing blame on each other, I feel it's really, it makes your mood worse*

knowing that, hey, I feel like I'm not the one doing bad here. Why are you blaming me?" To resolve such conflicts and prevent victims from tilting, they suggested showing supportive messages based on the gameplay data: "*I think you could probably use AI as a way to encourage that player. Maybe the AI could analyze what that player did and see how that helped the overall team. For example, even if it's not that much, the AI could focus on that and tell the person, hey you did well in such and such*" (FG3P4). They also suggested pointing out to the blamer that they also have some room for improvement, and maybe they should not "*be the one talking*" (FG3P2). While some acknowledged that highlighting someone's mistakes is questionable (FG2P4, FG2P2), others felt it would be justified: "*if this person is lashing out at people saying you have awful damage, you have, why are you not doing any healing? And then the AI tells you that you've been on point for two seconds, this match, then that's kind of a way to be. It's forcing players to face themselves rather than just lashing out at everyone*" (FG3P3). They also noted that if not through the in-game agent, they would still look up player stats to decide how to respond: "*I would feel really good if someone was just reminding me the stats of the player. Because sometimes, before any player rages, the first thing they want to do is check the stats of the person that's going after them and if they're bad it's like, okay, don't even talk to me. But if they're doing really good it's like I feel more inclined to take the advice*". These views further highlight how preoccupied players were with skill and performance. Coming from an experienced player, even insults may be dismissed as advice.

As a final solution if nothing else worked, participants suggested nudging players to take a break: pause the game for a minute, change to casual game mode, log off for a while, or abort the game and let people leave without penalties. Pausing the game seemed a good idea at first, to deal with heightened emotions; it was, however, rejected for disrupting the game and prioritizing one player's needs above the team: "*those pauses or that interruption of the game affects other people as well. Especially when players are doing fine*" (FG1P4). Instead, they proposed using AI for recognizing fatigue and suggest (a) changing from competitive

to casual game mode ("*I think it could be a suggestion be like how about you try casual matches with a different range of skill sets?*" (FG2P4)) or (b) taking a break from the game: "*if AI is able to just kind of see your gameplay and maybe your chat and detect, oh, this person's having an emotional peak, I think having maybe one warning where it's like maybe it's time to take a break*" (FG3P7). To make this cue more convincing and strategic, others suggested showing how the performance declined: "*what if they gave you a little dose of reality? It's like you've been playing this game for this many hours, you've lost this many games*" (FG5P4), "*Just pack it up bro*" (FG5P5). Participants briefly discussed the option of aborting the game without penalties in case everyone was toxic: "*If the match is going nowhere and maybe it might just be best to offer a cancel match with no penalties and all that*" (FG1P3). Although the idea was welcomed by others, they also felt it was too easy to abuse "*if we're just losing really bad we could just all quit and then there's no penalty*" (FG1P4).

Although these supportive solutions were more accepted, participants pointed out that they could also be distracting and even become the source of toxicity. Reputation could be used to target players (FG4P6), and encouraging or discouraging messages can irritate them further (FG3P6). Many of them felt that they would be annoyed or even angered if an AI told them how to play or when to take a break: "*I think it would just get people peeved enough there saying, why are you trying to tell me what to do? And they're just going to turn it off and just say even worse things because they got a little bit more angry*" (FG5P5). Participants presumed that AI suggestions would hurt experienced players' egos, even if they might be helpful in lower ranks. As FG3P6 explained: "*I'm not top 1%, but I am very pretty top 3 and 10%* [player in Valorant]. [...] *And I think it would tilt me even more if I were to just get told what to do because in my mind I already know what to do. But yeah, so in more casual games or game mode or say even lower rank, anything below 50% tire, I think that would be helpful*".

Distraction was raised as another concern; participants argued that these tools must be optional (FG1P2) and AI human interaction needs to be implemented with a variety of voice/text lines, "*otherwise it becomes very annoying*" (FG5P2). It also has to be carefully tailored to the specific genre and game, ensuring it does not disrupt the player experience. For example, in fast-paced competitive games "*if there's a pop-up on the side or something that's like: Hey, try to stay calm while you're playing, then it's like that's heavily distracting or that can even block the screen that you need to be looking at*" (FG1P5). Voice lines can be equally disruptive, as FG1P5 clarified: "*if you have an audio cue, it could play over footsteps, and then you don't hear the footsteps, and then that's crucial information that you're missing*". Therefore, both pop-ups and voice lines should be positioned and timed thoughtfully (FG2P2). The design ideation led participants to the concept of integrating the AI ally system into the game itself, allowing for a more natural and contextual support.

5.4.3.2 Supportive moderation integrated within the game

The concept of using existing characters in the game as AI allies emerged in many groups and was by far the most cherished idea. While participants had reservations about an external AI tool, seeing it as annoying and disruptive, they welcomed the same type of help from a beloved game character. "*So you know how some players have a very strong attraction or appeal to certain characters.* [...] *Say like you were in a heated game, right? The world is coming to an end, and your favorite character is telling you, Hey man, I think you should take a break.* [...] *I think it has an effect on you*" (FG3P4). FG1P4 voiced a similar sentiment: "*I play Marvel Rivals and if I heard Spiderman saying, chill out, you bet I'm chilling out*".

Ideas on how to implement this feature ranged from using non-diegetic customized NPCs to breaking the fourth wall between player and character. One group argued that it has to be non-diegetic, because "*it would be too off-putting if the character in the game started interacting with you in real time*" (FG2P2). Every other group felt that the deeper integration would increase the impact and found it humorous or shocking to interact with in-game

characters. They proposed using either existing NPCs or the player character as a supporting agent. For example, "*If all of a sudden your character breaks the fourth wall, maybe if there's something, a part of it that would immediately give you a shock factor, it would probably make you laugh, which immediately would change your mood.*" (FG1P5).

Participants found this solution ideal for narrative games or hero games (Shooter, MOBA) featuring characters with strong backgrounds and complex personalities. "*I'll say specifically that I know Overwatch and Marvel Rivals, those games are, what is it? Like hero games, ... like their draw are the characters*" (FG4P2). These existing characters were seen as ideal candidates to be trusted allies: they were already part of the game world, and their personalities and backgrounds made them relatable and appealing to the players. "*I think that the characters already have their own personalities, their own background that the players connect with, so they know them* [...] *that it's more personal*" (FG3P4). When they interact with the player, they can speak in their own voices in a style matching their personalities. For example, "*The important thing is the voice has to have a place. The character has a place in your game. If you're playing Rainbow Six Siege or something, then there could be a commander voice who's in your team and intervening, which is realistic to that extent, like clearing comms basically in real life*" (FG3P3).

In terms of how the AI ally could be helpful, participants revisited previous ideas and brainstormed new ones, including providing emotional support through de-escalating anger through humor, encouraging messages, nudging players to relax or take a break, and even working as chat moderators. Participants thought that their favored characters might help them with impulse control when they were tilting. Some felt they would only need a reminder "*that in the end it's just a game and it doesn't need to be shocking or anything. It just needs to be, hey, it's not that serious.*" (FG1P3), while others preferred humorous distraction: "*It doesn't even have to be 'calm down' or 'take a step away from the game', but it has to* [be], *I guess, something that's, I guess, kind of funny or lighthearted*" (FG3P2). Others suggested

positive encouragement and spotlighting strengths (FG4P2). For example: "*my first game of the day, this person or my team was just really bad. And then your coach at the end is saying, Hey, you did everything best as you could. Then you just go next. It's like, okay, I don't feel as bad*" (FG5P2).

Interestingly, participants generally vetoed open AI mediation between team members. They saw such public interventions as disruptive and, more importantly, as potential sources of escalating toxic behavior. They argued that any public exposure would single out the victim or the perpetrator and would potentially lead to more targeted attacks, because "*now this person's the laughing stock*" (FG5P2). Instead of using AI to openly support victims or call out perpetrators, they suggested incentivizing players to help those who struggled. An example was adding quests to support a less skilled teammate: "*It could also help improve the gameplay as well and maybe give an honor system to people who actually do an in-game quest of: 'Help Hello Kitty, survive two minutes', get a hundred percent exp or something for next round. I think it could also be a good implementation to improve gameplay* [...] *without affecting too much of that actual gameplay*" (FG3P4). Another suggestion was to encourage bystanders to step up through a reward system: "*Maybe you could have a system where maybe if someone is working to de-escalate the situation, calm the other person down, you could reward that person.*" (FG2P2).

The only public AI intervention participants supported was using the player characters to moderate the text chat. Instead of filtering or blocking hurtful messages, they suggested that the character controlled by the toxic player could rephrase the insult to be more neutral: "*Say I was playing, I don't know, Jax*[1]*, and it'd be like I type some really mean message to my mid laner and then the chat would appear to be like a Jax is saying, Hey, top-laner is not happy right now.* [...] *a way to make it clear that you're still mad at that person, but you're unable to get across the words that make the message extremely demeaning.*" (FG1P7).

[1]Jax is a top-laner dualist champion in League of Legends. https://www.leagueoflegends.com/en-us/champions/jax/

Participants agreed that every player is different and may need different forms of support to effectively self-regulate. Therefore, they suggested the AI ally be customizable: "*for someone like you who likes it to just take a step back, the AI would choose to maybe say a message like 'calm down, maybe take a breather'. And then maybe for players, like how you said, if you want something to distract them, you could play music or tell a joke or something like that, which I think is also great. All these ideas are great ideas. We're all different people.*" (FG4P2). Aside from setting preferred modes of support, they also urged personalization in the delivery style (FG5P5) that, for example, could be "*playful or, you know, those crazy drill sergeants from the military*" (FG5P2).

Another important factor was the ability to turn the feature off: "*Would there be an option to where you consciously say, if I'm playing really well in a game, I'm in this absolute flow state and I don't want to hear positive reinforcement, I don't want to be distracted, I want laser focus on what's going on in front of me*" (FG3P3). Or, as FG5P5 put it, "*It should be toggleable. Well, if someone wants to listen to it, they can if they don't want to, it's all right*". They pointed out that, with time, "*it's going to get boring or it's going to get stale, where it's like, okay, we're used to it*" (FG1P4), and suggested that the developers should frequently update this feature (FG1P4).

5.5 Discussion and Future Work

5.5.1 Attitudes toward AI moderation

Attitudes toward AI moderation systems were characterized by skepticism and distrust rooted in experiences of unfair decisions, fears of losing human agency and oversight, and not knowing how AI is used in moderation. Echoing previous work [216], well-known issues of fairness, human agency, privacy, and data protection emerged in the conversations. All of these were connected to the overarching problem of transparency. Although this list is not

exhaustive, they are the most salient concerns players shared in this study and in previous work [216].

Issues that impacted player experience, such as fairness and reliability, were at the center of the conversation. Privacy and data protection were less prominent topics, while issues, such as algorithmic bias or inclusion, were absent from the discussions. Social exclusion due to algorithmic bias is a long-standing issue [109, 110], one that poses a significant risk of further increasing the already prevalent discrimination against marginalized players in multiplayer online games [216].

Regarding fairness, participants recounted several examples of unfair decisions due to the perceived limitations of AI systems, specifically, the lack of human judgment and contextual understanding. Related to fairness, they stressed the importance of human agency and control, expressing the need for human oversight in the moderation process. Recognizing the problem of scalability, they urged the use of AI for menial tasks, such as flagging harmful content, and leaving the final decision to humans with a nuanced understanding and empathy. Previous work highlighted the importance of user-human-AI collaboration in AI-assisted moderation [202] and the benefits of community involvement in moderation and players' capacity to participate in moderating their communities [127, 133, 228, 229].

Privacy and data protection concerns were less prevalent, and participants had varying attitudes. While many were wary of constant surveillance, excessive data collection, and data sharing with third parties, some were less concerned about privacy and data practices, including using player data to train AI algorithms, as long as it wasn't misused.

Both fairness and privacy-related conversations revealed the underlying problem of transparency, or the lack of it. While discussing the fairness of AI moderation, participants often expressed uncertainty about whether AI was used at all in moderation, and if it was, how it worked. For example, while talking about AI moderation, they were unsure whether the

games they played used dictionary-based automated tools or actual AI, such as machine learning or natural language processing in chat moderation. Similarly, they appeared largely unaware of what data was collected from them, how it was stored or managed, who had access to it, and for what purposes. They acknowledged that privacy policies may contain some of this information, but admitted to not reading them. Research has found that policy contracts, including privacy policies, are often obscure legal texts produced to protect company interests rather than to inform players [23, 238] Transparency is a fundamental part of trust and safety, which should start with clearly communicated policies and practices regarding privacy, data governance, and the moderation infrastructure. Privacy and data management can be made more accessible through in-game notifications and user-friendly privacy settings [216]. Communicating the moderation process is just as important; at a minimum, players need to know what parts are automated, to what extent, and how human moderators and other stakeholders are involved. Research has also shown the importance of transparency and explainability in automated decision-making [29, 88, 128, 159, 160]. Players' lack of understanding of the AI systems governing their experiences is especially problematic, given that these technologies have further amplified top-down retributive moderation, raising significant ethical implications for players' privacy and autonomy [168, 216]. The potential use of explainable AI (XAI) [54], "white boxing", and interactive explanation [30] of AI systems may help improve transparency and player understanding [216]

Transparency in how these algorithms are designed and how they work is equally crucial. Governing algorithms codify critical social values and require normative decisions [151]. Such decisions are routinely made with the exclusion of the largest group of stakeholders, the players, reinforcing unilateral power structures.

5.5.2 Acceptable and helpful real-time AI interventions

5.5.2.1 Questioning punitive interventions

Participants in each focus group envisioned several creative ways in which AI could be leveraged in real-time moderation. Many were based on existing industry tools, and many involved novel interventions. Without having a "frame of reference", participants struggled to think beyond the boundaries of punitive systems they were so accustomed to. Consequently, exploring punitive solutions dominated the first part of the ideation. This echoes prior work that highlights players' entrenched assumptions about retributive approaches [76, 258]. Even when prompted to think of ways AI could be supportive, retributive solutions kept reappearing, showing the strong influence of the punitive baseline industry moderation.

Penalizing suggestions were based on standard moderation tools, such as censorship, restricting chat privileges, and banning players. Additionally, participants discussed gradually escalating punishments and restoring privileges, segregating players based on their behavioral data, and replacing toxic players with bots. All of these solutions are currently implemented in multiplayer online games today, with restricting participation as the default disciplinary measure. While participants were not against punishments and emphasized that suffering the consequences is necessary to correct behavior, they rejected most of the punitive solutions. Even with the perfect AI in mind, and despite the assumed "fairness", these solutions were unacceptable because they did not align with participants' core values. Censorship was found to be invasive, especially when it rephrased the toxic message instead of blocking it. They became cautious about replacing human players with bots, even if those might be "*more valuable than a griefing player*" (FG1P7). Playing with and against AI in a social game did not feel right, and the deceptive way companies currently use AI substitutes to keep players engaged by "*sprinkling in wins and losses*" (FG5P4) was seen as questionable. Segregating players also started as a possible solution, but was rejected for causing harm by amplifying and normalizing toxicity for groups of players.

5.5.2.2 Reimagining supportive AI - Integrated moderation

After scrutinizing several punitive options, participants gradually turned toward supportive solutions. Acknowledging that they themselves were often toxic to others and shifting the focus to self-regulation and supporting players, including perpetrators, helped them move away from the default punitive approach. Participants envisioned many forms of real-time contextual support based on behavioral and performance data. Some ideas were again based on existing solutions, such as using reputation systems to track and display players' social behavior. Other suggestions were less prevalent in current games, such as offering just-in-time guidance or nudging players with encouraging and discouraging messages to self-regulate when tilting and resist toxic outbursts.

Instead of moral alignment, acceptance was now a question of the applicability and practicality of these supporting tools in existing games. Despite their apparent value, tips and nudges were dismissed as unsolicited advice, frustrating players and undermining competence and autonomy. In addition to potential resistance from players, another concern was that text pop-ups and voice lines that risk suppressing critical visual and audio cues may disrupt the player experience. The idea of integrating supportive interventions with the game mechanics emerged as a natural solution to these issues, reflecting previous findings on novel forms of AI moderation [76]. Game integration dramatically increased acceptance and also slightly changed the proposed interventions. Instead of relying solely on gameplay data, solutions became more focused on self-regulation, empathy, and perspective-taking. Participants across groups were genuinely enthusiastic about the opportunity of getting real-time support from beloved game characters. As opposed to external AI tools that are both out of place and obtrusive, game characters are already part of the game world, with backgrounds and personalities designed to be believable and foster emotional attachment [14, 143, 184]. Bopp et al. [14] found that players develop different forms of attachment to player characters and NPCs with distinct emotional qualities [14]. Emotional bonds may form even when

players value characters for their utility [156], or view them as instruments through which they can express their playstyle [66]. Such findings also highlight the potential of leveraging players' affection for characters in supportive moderation. Instead of being realized as an external voice of authority, a diegetic integration can turn supportive AI interventions into an authentic and meaningful part of play [197].

While punitive interventions cannot be optional or personalized, supportive solutions were imagined with these options in mind. In agreement with prior work [76, 193], participants expressed a strong desire for control and agency. They emphasized the importance of being able to turn these features off and to customize them to meet individual preferences and needs. The large variety of characters in competitive, team-based genres is a unique opportunity for customizing the AI ally. It allows for selecting a supportive character with a personality and style that would be most beneficial for the player, e.g., guiding with empathy, humor, or authority. As long as the nudges matched the character's personality and background and fit the game context, participants did not see them as disruptive.

AI support was generally seen as a private matter, and no public forms of addressing toxicity were acceptable, as public exposure risked singling out and humiliating players. Yet, while AI was seen as an inadequate "upstander", participants found it helpful in encouraging bystanders to step in.

Integrated AI allies were considered to be a promising opportunity to help them with emotion regulation, self-reflection, and perspective-taking, crucial components in cultivating healthy communities [187].

As a next step, research could explore how such integrated nudges might work and what ethical considerations might arise. Further, in agreement with previous work, I urge that players be included in designing innovative moderation systems. In today's fast-paced technological evolution, where AI is deployed rapidly in every aspect of everyday life, people are

seldom included in design decisions or have an opportunity to opt out of being subjected to automated systems. Including players in designing AI moderation tools provides a unique opportunity to overcome many issues plaguing current approaches, create better governance systems, and build healthier online communities. I am hopeful that player involvement may help develop ethical AI moderation practices, empower players, and produce a system that reflects their values and needs.

5.6 Limitations

This study is not without limitations. First, the participants were local to my home institution, based in North America. Their views and attitudes may not generalize to the broader population, especially considering the small sample size. Second, the discussion overwhelmingly focused on team-based and competitive games, such as MOBAs, Shooters, and Battle Royales, with most mentioned titles developed in the USA by large AAA companies. While findings may generalize to multiplayer games with stronger social ties, such as MMOs or even VR games, they might not be sufficiently nuanced or specific for individual games. Technical feasibility was not considered during the focus groups, and players imagined a "perfect AI" to help creative ideation. Further scrutiny of these ideas and the consideration of potential ethical issues are required. Exploring how supportive AI interventions could be integrated into specific titles could be beneficial. Future work could also explore how such interventions impact player experience through prototyping and experimental methods.

Chapter 6

Conclusion and Future Direction

In this chapter, I synthesize the findings from the three included studies in light of my thesis statements and discuss implications for the social safety and moderation apparatus.

Toxicity is a complex sociotechnical issue that cannot be solved by technological tools only; technological governance and moderation, however, play a crucial role in setting norms and fostering thriving communities. As participants in the focus group study emphasized, the community starts with game design. Narrative content and visual elements influence which communities form, and who feels the game was made for them. Design can challenge or reinforce systemic issues such as racism and sexism [68, 140, 154, 155, 203]. The game rules embedded in the code and meta-systems greatly influence how players perceive and treat each other. Weak sociality, ephemeral teams, competition between team members, and valorizing performance and rank over sportsmanship may all contribute to players' reduced perspective-taking, lack of empathy, and seeing one another as instruments [124, 131, 134, 206, 236]. Some of these designs may be deliberate dark patterns (e.g., manipulative features to increase engagement) [135], but many are unintentional. Fast-paced competitive play, for example, also increases stress and often leads to friction between players [2, 82, 124, 206, 234]. Due to

developmental characteristics, teens and young adults, the primary audiences of social and competitive games, are especially vulnerable to toxic behavior in these high-stress environments [49, 97]. Balancing potential unintended consequences for player dynamics without compromising the desired player experience is not easy, yet it is crucial for fostering thriving communities. While careful design may alleviate some grievances and reduce toxicity, social play will always be prone to friction between players.

Technologies of player safety, privacy, and moderation – the main interest of this thesis – are the other half of technological governance, aiming to protect players, prevent and address harm.

The findings of the three included studies echo prior work emphasizing the shortcomings of the common punitive moderation paradigm [129, 130, 132, 258] and calling for a shift towards restorative justice [258], care, and supportive practices [76, 187, 216]. In this dissertation, I demonstrate that current safety controls and moderation systems cannot effectively protect players and are not sufficiently aligned with the needs of the broad playerbase.

Both moderation and responding to toxicity have a norm-setting effect. Leaving toxicity unchallenged and participating in its perpetuation contribute to the normalization of such behavior [11, 120, 206]. Results of Chapter 3 show that responding to toxicity is habitual and driven by prior game experiences and toxic exposure. While cumulative bystander experience leads to more social withdrawal and may reinforce the bystander effect, victimization and game experience trigger active coping: retaliation, or productively confronting toxicity. This suggests that with toxic exposure and game involvement, players not only internalize toxic norms but also develop resilience and productive coping strategies. Calling out perpetrators or lashing out at them are not far apart. Supporting players in standing up against toxicity productively can be invaluable in breaking the cycle of harm and cultivating thriving communities. Players' views on the effectiveness of moderation tools align with their preferred response strategies. Although current proactive tools (EULA, Code of Conduct, priming,

and reputation systems) were rated the least effective, those who more often respond productively or log off rather than endure toxicity value these tools more. Supporting them with **improved proactive measures may help reinforce productive confrontation and reduce player churn. Such tools may also be valuable in encouraging bystanders, who currently prefer social withdrawal when facing or witnessing toxicity.** Additionally, punitive measures and protective tools that players considered more valuable do not address the root causes of toxicity. Aside from profanity filtering, all of these are reactive interventions that happen after the fact, and none attempt to restore relationships or provide care for victims.

Chapter 4 confirms and extends these findings. **Both parental controls and in-game social safety settings support restrictions and limiting participation, while missing out on opportunities such as encouraging communication between parent and teen, educating teens about online social safety through suggested settings and clear explanations, or allowing for scaffolded autonomy and co-setting rules.** With a few exceptions (e.g., EA), parental controls on platforms with 13+ audiences follow the same restrictive practices as those designed for children, infantilizing teens and prioritizing parental authority over teens' developmental needs for increased autonomy and privacy. Monitoring, for example, is of value for both parental mediation and teen self-regulation. For supporting children and tweens, a shared dashboard of playtime, purchases, and other game activities could be beneficial, providing opportunities for communication, negotiation, and self-reflection. As teens mature, privacy and the ability to self-regulate become increasingly important. For them, reduced surveillance and detailed private dashboards may be more suitable. Additional forms of active mediation are also much needed. Crucially, they should move beyond screen time and purchase requests and involve parents in discussions and interactions that help foster online safety and privacy without limiting engagement [251]. Importantly, I found that the boundaries of parental controls are often unclear. While platform-level spending and screentime limits usually apply to games, social

interaction settings don't always do. While this allows teens to self-regulate, neither party is sufficiently informed.

Unlike parental controls, in-game safety settings are not grouped in a dedicated menu; they fail to present a coherent system and leave it to players to discover all of the related controls. The low granularity and lack of customization and transparency also show how neglected these features are. There are plenty of opportunities for improvement; it could start with having a dedicated place for social safety and privacy, adequate information on how to control them, friends-only defaults, and reminders for players to review and set their preferences. In contrast, reactive tools, such as muting, blocking, and reporting, are a lot more accessible. Muting shows the most careful design, being easily available, unambiguous, and granular enough for players to customize communication by lowering individual players' audio and muting different modalities (voice, text, or pings) or channels (team, party, or an individual player) of communication. Muting is highly usable and grants players autonomy and control to take immediate, adequate action that best suits their needs in the given situation. Making this feature the most used and appreciated tool in the safety and moderation apparatus (see, e.g., Chapter 3 and Refs. [235] and [213]). **Crucially, however, no tools are offered for real-time active coping or impulse control. Such interventions would not only be helpful for teens to practice self-regulation but would benefit players in general, especially in competitive and team-based games where stress, performance anxiety, and friction between players are common [2, 124, 206, 207, 235].**

What should such tools be like? Could AI be leveraged in providing real-time support for players to de-escalate toxicity? Chapter 5 explores these questions. Players are generally skeptical about AI moderation; their wariness is rooted both in their negative experiences with the punitive system and in the lack of transparency that leaves them in the dark about what AI is used for, how it works, what data is collected from them, and how it is used. This reflects previous work on players' unawareness of AI systems [216] and privacy prac-

tices [15] and aligns with critiques of the punitive automated moderation and its lack of transparency [125, 129, 159, 160]. Participants envisioned various punitive and supportive solutions and discussed different acceptance criteria, including potential ethical issues, unwanted consequences (e.g., further toxicity and agitation), and disrupting play and immersion. The initial ideas were tied to current punitive solutions, such as segregating players or replacing them with NPCs. These ideas, however, were all rejected for ethical reasons. Replacing humans with bots and creating toxic echo chambers didn't align with players' core values and were unacceptable to them. Interestingly, these are existing industry solutions. Among the supportive AI moderation ideas were, for example, nudges based on match data and behavioral profiles, encouraging self-reflection, perspective taking, and recognizing contribution. These solutions were more aligned with participants' values; however, they were considered unhelpful in the context of fast-paced competitive gaming. AI assistance was seen as an external, unsolicited help, another source of potential annoyance and frustration that disrupts play. **As a resolution, participants concluded that the most acceptable and helpful mode of real-time support would be to integrate the intervention into the game world.** Strong emotional attachment to game characters [14] and waiving the nudges into the narrative changed participants' affective valence from negative emotions, such as anger or frustration, to positive excitement and thrill. **Through diegetic integration, potential supportive AI interventions can become an authentic, meaningful part of play [197], rather than an external voice of authority. Additionally, supportive interventions – unlike punishment – can be optional and customizable, supporting players' need for autonomy and control over their game experience.**

Taken together, my research highlights the opportunity to expand the current moderation apparatus with proactive tools, such as educating players about AI tools and data collection practices, providing in-game social safety and privacy resources, and implementing in-game supportive interventions. Existing tools greatly miss out on norm-setting. The biggest challenge of moderation and one of its most important tasks is to create strong shared norms

within a community, both directly by explaining why a player moderated, and indirectly through spotlighting good behavior, fostering a sense of belonging, and welcoming newcomers [83]. Promoting self-regulation, including self-monitoring and real-time impulse control, appears particularly promising for teens and young adults. Unlike punitive tools, such interventions can be optional, customizable, contextual, and fully integrated with the game world, accommodating the players' need for autonomy, control, and immersive play. Leveraging players' emotional attachment to in-game characters presents an auspicious avenue for nudging prosocial behavior, perspective taking, and self-reflection.

6.1 Future Directions

Moving forward, I aim to continue investigating social safety and moderation in online games to support players and cultivate healthy communities. There are several directions I'm interested in, such as interviewing different stakeholders and refining and testing supportive interventions.

Family safety practices in online games. After assessing parental controls and in-game safety mechanisms, a natural next step would be exploring family safety practices through in-depth interviews with adolescents and their parents. I'm curious to learn about teens' experiences with parental controls and in-game social safety settings. How do they utilize these technologies and why? At the same time, I also plan to investigate parents' experiences. Learning how they perceive and use parental controls on gaming platforms and what additional strategies they have would provide valuable insights into families' gaming-related safety practices.

The future of moderation according to industry professionals. I also plan to interview industry professionals working in trust and safety teams, including product managers, policy experts, designers, moderators, and community managers. Their insights would be invaluable regarding current industry practices. Especially considering the recent shift in moderation,

such as Roblox's new aggressive AI-driven system. I'm excited to learn how industry professionals balance safety and player experience amid legal, ethical, technological, and other constraints. What are the biggest challenges they identify? And how do they see the future of moderation?

Participatory design workshops for supportive moderation interventions. Continuing my work on supportive interventions, I aim to conduct participatory design workshops with players to refine the initial ideas that the focus group participants envisioned. Many of whom expressed interest in returning for longer, engaged design sessions. I also plan to prototype and evaluate the most promising designs.

Bibliography

[1] L. Achterbosch, C. Miller, and P. Vamplew. A taxonomy of griefer type by motivation in massively multiplayer online role-playing games. *BEHAVIOUR & INFORMATION TECHNOLOGY*, 36(8):846–860, 2017. doi: 10.1080/0144929X.2017.1306109.

[2] S. Adinolf and S. Turkay. Toxic behaviors in esports games: Player perceptions and coping strategies. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts, pages 365–372, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359689. doi: 10.1145/3270316.3271545. URL `https://doi.org/10.1145/3270316.3271545`.

[3] M. AI. The impact of ai voice moderation on the call of duty player experience. https://www.modulate.ai/case-studies/modulate-activision-case-study, 2025.

[4] G. W. Allport. *The Nature of Prejudice*. Cambridge: Addison-Wesley Publishing Company, 1954.

[5] A. Alsoubai, J. Song, A. Razi, N. Naher, M. De Choudhury, and P. J. Wisniewski. From 'friends with benefits' to 'sextortion:' a nuanced investigation of adolescents' online sexual risk experiences. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), Nov. 2022. doi: 10.1145/3555136. URL `https://doi.org/10.1145/3555136`.

[6] Anti-Defamation League. Caught in a vicious cycle: Obstacles and opportunities for trust and safety teams in the games industry. https://www.adl.org/resources/report/caught-vicious-cycle-obstacles-and-opportunities-trust-and-safety-teams-games, 2022.

[7] Anti-Defamation League. Hate is no game: Hate and harassment in online games 2023. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023, 2023.

[8] A. Arango, J. Pérez, and B. Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331262. URL `https://doi.org/10.1145/3331184.3331262`.

[9] M. E. Ballard and K. M. Welch. Virtual warfare: Cyberbullying and cybervictimization in mmog play. *Games and Culture*, 12(5):466–491, 2017. doi: 10.1177/1555412015592473. URL https://doi.org/10.1177/1555412015592473.

[10] D. Baumrind. A developmental perspective on adolescent risk taking in contemporary america. *New Directions for Child and Adolescent Development*, 1987(37):93–125, 1987. doi: https://doi.org/10.1002/cd.23219873706. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cd.23219873706.

[11] N. A. Beres, J. Frommel, E. Reid, R. L. Mandryk, and M. Klarkowski. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445157. URL https://doi.org/10.1145/3411764.3445157.

[12] N. A. Beres, M. Klarkowski, and R. L. Mandryk. Playing with emotions: A systematic review examining emotions and emotion regulation in esports performance. *Proc. ACM Hum.-Comput. Interact.*, 7(CHI PLAY), Oct. 2023. doi: 10.1145/3611041. URL https://doi.org/10.1145/3611041.

[13] J. Blackburn and H. Kwak. Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 877–888, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2567987. URL https://doi.org/10.1145/2566486.2567987.

[14] J. A. Bopp, L. J. Müller, L. F. Aeschbach, K. Opwis, and E. D. Mekler. Exploring emotional attachment to game characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, page 313–324, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366885. doi: 10.1145/3311350.3347169. URL https://doi.org/10.1145/3311350.3347169.

[15] A. Bourdoucen, L. Nurgalieva, and J. Lindqvist. Privacy is the price: Player views and technical evaluation of data practices in online games. *Proc. ACM Hum.-Comput. Interact.*, 7(CHI PLAY), Oct. 2023. doi: 10.1145/3611064. URL https://doi.org/10.1145/3611064.

[16] N. Bowman and R. Kowert. *The Impact of Video Game Play on Human (and Orc) Creativity*, pages 41 – 62. Academic Press, 12 2015. ISBN 9780128014622. doi: 10.1016/B978-0-12-801462-2.00002-3.

[17] A. Braithwaite. Wowing alone: The evolution of "multiplayer"in world of warcraft. *Games and Culture*, 13(2):119–135, 2023/05/06 2015. doi: 10.1177/1555412015610246. URL https://doi.org/10.1177/1555412015610246.

[18] V. Braun and V. Clarke. *Thematic Analysis*, pages 1–7. Springer International Publishing, Cham, 2020. ISBN 978-3-319-69909-7. doi: 10.1007/978-3-319-69909-7_3470-2. URL https://doi.org/10.1007/978-3-319-69909-7_3470-2.

[19] V. Braun and V. Clarke. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health*, 24(1):1–6, 2023. doi: 10.1080/26895269.2022.2129597.

[20] T. A. Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[21] M. Brunner and H.-M. Suss. Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, 65(2):227–240, 2005. doi: 10.1177/0013164404268669. URL `https://doi.org/10.1177/0013164404268669`.

[22] D. Buckingham. *Youth, Identity, and Digital Media*. The MIT Press, 2007.

[23] T. Busch, K. Boudreau, and M. Consalvo. *Toxic Gamer Culture, Corporate Regulation, and Standards of Behavior among Players of Online Games*, pages 176–190. Routledge, 10 2015. ISBN 9781315748825. doi: 10.4324/9781315748825-13.

[24] A. Canossa, D. Salimov, A. Azadvar, C. Harteveld, and G. Yannakakis. For honor, for toxicity: Detecting toxic behavior through gameplay. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY), oct 2021. doi: 10.1145/3474680. URL `https://doi.org/10.1145/3474680`.

[25] C. Carver, M. Scheier, and J. Weintraub. Assessing coping strategies: A theoretically based approach. *Journal of personality and social psychology*, 56:267–83, 03 1989. doi: 10.1037//0022-3514.56.2.267.

[26] F.-C. Chang, C.-H. Chiu, N.-F. Miao, P.-H. Chen, C.-M. Lee, T.-F. Huang, and Y.-C. Pan. Online gaming and risks predict cyberbullying perpetration and victimization in adolescents. *INTERNATIONAL JOURNAL OF PUBLIC HEALTH*, 60(2):257–266, FEB 2015. doi: 10.1007/s00038-014-0643-x.

[27] V. H. H. Chen and G. S. Chng. Active and restrictive parental mediation over time: Effects on youths' self-regulatory competencies and impulsivity. *Computers & Education*, 98:206–212, 2016. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2016.03.012. URL `https://www.sciencedirect.com/science/article/pii/S0360131516300756`.

[28] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, page 71–80, USA, 2012. IEEE Computer Society. ISBN 9780769548487. doi: 10.1109/SocialCom-PASSAT.2012.55. URL `https://doi.org/10.1109/SocialCom-PASSAT.2012.55`.

[29] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for

Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300789. URL https://doi.org/10.1145/3290605.3300789.

[30] Z. Cheng, D. Cheng, Y. Yang, A. Cheng, C. Tan, and Y. Zhuang. What makes a good team? a large-scale study on the effect of team composition in honor of kings. page 2666 – 2672, 2019. doi: 10.1145/3308558.3313530. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066888735&doi=10.1145%2f3308558.3313530&partnerID=40&md5=b25e89c3d1ff9fee896f1470fb88140e. Cited by: 8; All Open Access, Green Open Access.

[31] T. Chesney, I. Coyne, B. Logan, and N. Madden. Griefing in virtual worlds: causes, casualties and coping strategies. *Information Systems Journal*, 19(6):525–548, 2009. doi: https://doi.org/10.1111/j.1365-2575.2009.00330.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.2009.00330.x.

[32] S. Choudhury, S.-J. Blakemore, and T. Charman. Social cognitive development during adolescence. *Social Cognitive and Affective Neuroscience*, 1(3):165–174, 12 2006. ISSN 1749-5016. doi: 10.1093/scan/nsl024. URL https://doi.org/10.1093/scan/nsl024.

[33] E. Conroy, M. Kowal, A. J. Toth, and M. J. Campbell. Boosting: Rank and skill deception in esports. *Entertainment Computing*, 36:100393, 2021. doi: https://doi.org/10.1016/j.entcom.2020.100393. URL https://www.sciencedirect.com/science/article/pii/S1875952120301014.

[34] M. Consalvo. Confronting toxic gamer culture: A challenge for feminist game studies scholars. *Ada: A Journal of Gender, New Media, and Technology*, 2012. URL http://adanewmedia.org/2012/11/issue1-consalvo.

[35] C. Cook. *Everything You Never Wanted to Know about Trolls: An Interdisplinary Exploration of the Who's, What's, and Why's of Trolling in Online Games*. PhD thesis, Tilburg University, 2021. Series: TiCC Ph.D. Series Volume: 72.

[36] C. Cook and R. Kowert. Toxic waste removal: Toxicity in games and the perceived effectiveness of reporting. In *2021 DiGRA International Conference*, 06 2021.

[37] C. Cook, J. Schaafsma, and M. Antheunis. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9):3323–3340, 2018. doi: 10.1177/1461444817748578. URL https://doi.org/10.1177/1461444817748578. PMID: 30581367.

[38] C. Cook, R. Conijn, J. Schaafsma, and M. Antheunis. For whom the gamer trolls: A study of trolling interactions in the online gaming context. *JOURNAL OF COMPUTER-MEDIATED COMMUNICATION*, 24(6):293–318, NOV 2019. ISSN 1083-6101. doi: 10.1093/jcmc/zmz014.

[39] C. L. Cook, S. Y.-C. Tang, and J.-H. T. Lin. Billy goats crossing the cyber-bridge: Interviews exploring the experiences, coping techniques, and intervention desires of in-game trolling targets. *COMMUNICATION MONOGRAPHS*, 2023 SEP 7 2023. ISSN 0363-7751. doi: 10.1080/03637751.2023.2252487.

[40] J. A. Cornel, C. Christian Pablo, J. A. Marzan, V. Julius Mercado, B. Fabito, R. Rodriguez, M. Octaviano, N. Oco, and A. D. La Cruz. Cyberbullying detection for online games chat logs using deep learning. In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5, 2019. doi: 10.1109/HNICEM48295.2019.9072811.

[41] A. C. Cote. "i can defend myself": Women's strategies for coping with harassment while gaming online. *GAMES AND CULTURE*, 12(2):136–155, MAR 2017. doi: 10.1177/1555412015587603.

[42] J. L. Cotler, M. Fryling, and J. Rivituso. Causes of cyberbullying in multi-player online gaming environments: Gamer perceptions. *Journal of Information Systems Applied Research*, 10:4–14, 2017. ISSN 1.

[43] K. Crawford and T. Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016. doi: 10.1177/1461444814543163. URL https://doi.org/10.1177/1461444814543163.

[44] N. Crenshaw. Social experience in world of warcraft: Technological and ideological mediations. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY Companion '16, pages 1–4, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344586. doi: 10.1145/2968120.2990468. URL https://doi.org/10.1145/2968120.2990468.

[45] N. Crenshaw and B. Nardi. "it was more than just the game, it was the community": Social affordances in online games. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 3781–3790, 2016. ISBN 1530-1605. doi: 10.1109/HICSS.2016.471.

[46] N. Crenshaw and B. Nardi. Npcs as social mediators in massively multiplayer online games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 11(4):89–91, 2023/05/07 2021. doi: 10.1609/aiide.v11i4.12829. URL https://ojs.aaai.org/index.php/AIIDE/article/view/12829.

[47] M. Curcio. Cybercrime and video games: Exploring children's safety in the gaming environment. In *Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference*, EICC '23, page 144–148, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398299. doi: 10.1145/3590777.3590801. URL https://doi.org/10.1145/3590777.3590801.

[48] J. M. Cénat, M. Hébert, M. Blais, F. Lavoie, M. Guerrier, and D. Derivois. Cyberbullying, psychological distress and self-esteem among youth in quebec schools. *Journal of Affective Disorders*, 169:7–9, 2014. ISSN 0165-0327. doi: https://doi.org/10.1016/j.jad.2014.07.019. URL https://www.sciencedirect.com/science/article/pii/S0165032714004509.

[49] R. E. Dahl. Adolescent brain development: A period of vulnerabilities and opportunities. keynote address. *Annals of the New York Academy of Sciences*, 1021(1): 1–22, 2004. doi: https://doi.org/10.1196/annals.1308.001. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1308.001.

[50] G. Dale, F. Kattner, D. Bavelier, and C. Green. Cognitive abilities of action video game and role-playing video game players: Data from a massive open online course. *Psychology of Popular Media*, 9:347–358, 05 2019. doi: 10.1037/ppm0000237.

[51] Dice. Breaking down video gamer demographics. https://www.dice.com/career-advice/breaking-down-video-games-demographics, 2025.

[52] L. D. Dorn, R. E. Dahl, H. R. Woodward, and F. Biro. Defining the boundaries of early adolescence: A user's guide to assessing pubertal status and pubertal timing in research with adolescents. *Applied Developmental Science*, 10(1):30–56, 2006. doi: 10.1207/s1532480xads1001_3. URL https://doi.org/10.1207/s1532480xads1001_3.

[53] C. Drejer, S. S. Sabet, G. A. Baugerud, and M. A. Riegler. It's all in the game-an exploration of extensive communication on gaming platforms and the risks of online sexual grooming. *Available at SSRN 4671140*, 2024.

[54] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), Jan. 2023. ISSN 0360-0300. doi: 10.1145/3561048. URL https://doi.org/10.1145/3561048.

[55] M. S. Eastin, B. S. Greenberg, and L. Hofschire. Parenting the internet. *Journal of Communication*, 56(3):486–504, 09 2006. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2006.00297.x. URL https://doi.org/10.1111/j.1460-2466.2006.00297.x.

[56] N. Eisenberg and A. S. Morris. Moral cognitions and prosocial responding in adolescence. *Handbook of adolescent psychology*, pages 155–188, 2004.

[57] P. R. C. Emily A. Vogels. The state of online harassment. Web page, 2021. URL https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

[58] Entertainment Software Association. 2022 essential facts about the video games industry. https://www.theesa.com/wp-content/uploads/2022/06/2022-Essential-Facts-About-the-Video-Game-Industry.pdf, 2022.

[59] Entertainment Software Association. 2023 essential facts about the us video games industry. $https://www.theesa.com/wp-content/uploads/2023/07/ESA_2023_Essential_Facts_FINAL_07092023.pdf$, 2023.

[60] Entertainment Software Association. 2024 essential facts about the u.s. video game industry. https://www.theesa.com/wp-content/uploads/2024/05/Essential-Facts-2024-FINAL.pdf, 2024.

[61] J. Fahlström and E. C. Matson. Preventing toxic behaviour through game mechanics. 2014. URL http://uu.diva-portal.org/smash/get/diva2:731451/FULLTEXT01.pdf.

[62] Fair Play Alliance and Anti-Defamation League. Disruption and harms in online gaming framework. https://www.adl.org/resources/report/disruption-and-harms-online-gaming-framework, 2020.

[63] A. Faraz, J. Mounsef, A. Raza, and S. Willis. Child safety and protection in the online gaming ecosystem. *IEEE Access*, 10:115895–115913, 2022. doi: 10.1109/ACCESS.2022.3218415.

[64] D. Finkelhor, H. Turner, D. Colburn, and W. Walsh. Online child sexual exploitation and abuse, findings from a national survey. Crimes Against Children Research Center Bulletin, 2023. URL https://www.unh.edu/ccrc/sites/default/files/media/2023-11/ocsea-slide-bulletin-nov-5-2023.pdf. Retrieved June 26, 2025.

[65] C. Y. Foo and E. M. I. Koivisto. Defining grief play in mmorpgs: Player and developer perceptions. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, volume 74 of *ACE '04*, pages 245–250, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138822. doi: 10.1145/1067343.1067375. URL https://doi.org/10.1145/1067343.1067375.

[66] C. M. Ford. Virtuosos on the screen: Playing virtual characters like instruments in competitive super smash bros. melee. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1935–1948, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3026053. URL https://doi.org/10.1145/3025453.3026053.

[67] F. O. S. I. (FOSI). Connected and protected: Insights from fosi's 2025 online safety survey. Web page, 2025. URL https://fosi.org/research/connected-and-protected-insights-from-fosis-2025-online-safety-survey/.

[68] J. Fox and W. Y. Tang. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Computers in Human Behavior*, 33:314–320, 2014. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2013.07.014. URL https://www.sciencedirect.com/science/article/pii/S0747563213002525.

[69] J. Fox and W. Y. Tang. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society*, 19(8):1290–1307, 2017. doi: 10.1177/1461444816635778. URL https://doi.org/10.1177/1461444816635778.

[70] J. Fox, M. Gilbert, and W. Y. Tang. Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media & Society*, 20(11):4056–4073, 2018. doi: 10.1177/1461444818767102. URL https://doi.org/10.1177/1461444818767102.

[71] A. Frik, J. Kim, J. R. Sanchez, and J. Ma. Users' expectations about and use of smartphone privacy and security settings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517504. URL https://doi.org/10.1145/3491102.3517504.

[72] J. Frommel and R. L. Mandryk. Toxicity in online games: The prevalence and efficacy of coping strategies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642523. URL https://doi.org/10.1145/3613904.3642523.

[73] J. Frommel, D. Johnson, and R. L. Mandryk. How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness. *COMPUTERS IN HUMAN BEHAVIOR REPORTS*, 10, MAY 2023. doi: 10.1016/j.chbr.2023.100302.

[74] J. Frommel, R. Mandryk, and D. Johnson. The role of self-regulation and need satisfaction and frustration for predicting toxic behaviour in games. Preprint on ResearchGate, 01 2025.

[75] M. Fryling, J. L. Cotler, J. Rivituso, L. Mathews, and S. Pratico. Cyberbullying or normal game play? impact of age, gender, and experience on cyberbullying in multiplayer online gaming environments: Perceptions from one gaming forum. *Journal of Information Systems Applied Research*, 8:4–18, 2015. ISSN 1.

[76] R. Galwey, L. A. Sparrow, and M.-A. Butt. Beyond the enforcer: Exploring metaphors for alternative ai moderation in online games. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3720148. URL https://doi.org/10.1145/3706599.3720148.

[77] A. Giardina, M. Di Blasi, A. Schimmenti, D. King, V. Starcevic, and J. Billieux. Online gaming and prolonged self-isolation: Evidence from italian gamers during the covid-19 outbreak. *Clinical Neuropsychiatry*, 18:65–74, 02 2021. doi: 10.36131/cnfioritieditore20210106.

[78] T. Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. 01 2018. ISBN 9780300235029. doi: 10.12987/9780300235029.

[79] T. Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020. doi: 10.1177/2053951720943234. URL https://doi.org/10.1177/2053951720943234.

[80] R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*,

7(1):2053951719897945, 2020. doi: 10.1177/2053951719897945. URL https://doi.org/10.1177/2053951719897945.

[81] T. D. Grace, I. Larson, and K. Salen. Policies of misconduct: A content analysis of codes of conduct for online multiplayer games. *Proc. ACM Hum.-Comput. Interact.*, 6(CHI PLAY), Oct. 2022. doi: 10.1145/3549513. URL https://doi.org/10.1145/3549513.

[82] K. Grandprey-Shores, Y. He, K. L. Swanenburg, R. Kraut, and J. Riedl. The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, pages 1356–1365, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325400. doi: 10.1145/2531602.2531724. URL https://doi.org/10.1145/2531602.2531724.

[83] J. Grimmelmann. The virtues of moderation. *Yale Journal of Law and Technology*, 17:2, 2015. URL https://api.semanticscholar.org/CorpusID:8354427.

[84] T. Gupta and P. Gehlawat. Emotional regulation in adolescents: A narrative review. *Journal of Indian Association for Child and Adolescent Mental Health*, 16(3):171–193, 2020.

[85] M. Gámez-Guadix, I. Orue, P. K. Smith, and E. Calvete. Longitudinal and reciprocal relations of cyberbullying with depression, substance use, and problematic internet use among adolescents. *Journal of Adolescent Health*, 53(4):446–452, 2013. ISSN 1054-139X. doi: https://doi.org/10.1016/j.jadohealth.2013.03.030. URL https://www.sciencedirect.com/science/article/pii/S1054139X13002140.

[86] Y. J. Halbrook, A. T. O'Donnell, and R. M. Msetfi. When and how video games can be good: A review of the positive effects of video games on well-being. *Perspectives on Psychological Science*, 14(6):1096–1104, 2019. doi: 10.1177/1745691619863807. URL https://doi.org/10.1177/1745691619863807. PMID: 31672105.

[87] Y. J. Halbrook, A. T. O'Donnell, and R. M. Msetfi. When and how video games can be good: A review of the positive effects of video games on well-being. *Perspectives on Psychological Science*, 14(6):1096–1104, 2019. doi: 10.1177/1745691619863807. URL https://doi.org/10.1177/1745691619863807. PMID: 31672105.

[88] K. Haresamudram, S. Larsson, and F. Heintz. Three levels of ai transparency. *Computer*, 56(2):93–100, 2023. doi: 10.1109/MC.2022.3213181.

[89] H. Hartikainen, N. Iivari, and M. Kinnula. Should we design for control, trust or involvement? a discourses survey about children's online safety. In *Proceedings of the The 15th International Conference on Interaction Design and Children*, IDC '16, page 367–378, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343138. doi: 10.1145/2930674.2930680. URL https://doi.org/10.1145/2930674.2930680.

[90] J. Henseler, C. M. Ringle, and M. Sarstedt. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 1 2014. doi: 10.1007/s11747-014-0403-8. URL https://doi.org/10.1007/s11747-014-0403-8.

[91] Z. Hilvert-Bruce and J. T. Neill. I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior*, 102:303–311, 2020. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2019.09.003. URL https://www.sciencedirect.com/science/article/pii/S0747563219303310.

[92] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221, 2010. doi: 10.1080/13811118.2010.494133. URL https://doi.org/10.1080/13811118.2010.494133. PMID: 20658375.

[93] H. J. Hong and J. Connelly. High e-performance: esports players' coping skills and strategies. *International Journal of Esports*, 1, 2022. doi: https://doi.org/10.1016/j.chb.2019.09.003. URL https://www.ijesports.org/article/93/html.

[94] C. House of Representatives. 47 u.s.c. 230 - protection for private blocking and screening of offensive material. https://www.govinfo.gov/app/details/USCODE-2011-title47/USCODE-2011-title47-chap5-subchapII-partI-sec230/summary, 2011.

[95] L.-t. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[96] D. L. Hughes, A. A. Camden, and T. Yangchen. Rethinking and updating demographic questions: Guidance to improve descriptions of research samples. *Psi Chi Journal of Psychological Research*, 21(3):138–151, 2016. doi: 10.24839/2164-8204.JN21.3.138.

[97] N. Jaworska and G. MacQueen. Adolescence as a unique developmental period. *Journal of psychiatry & neuroscience: JPN*, 40(5):291, 2015.

[98] P. R. C. Jeffrey Gottfried and Olivia Sidoti. Teens and video games today. Web page, 2024. URL https://www.pewresearch.org/internet/2024/05/09/teens-and-video-games-today/.

[99] Jessica Clement. Distribution of video gamers in the united states in 2023, by generation. https://www.statista.com/statistics/189582/age-of-us-video-game-players/, 2024.

[100] Jessica Clement. Video gaming market in the united states - statistics & facts. https://www.statista.com/chart/22392/global-revenue-of-selected-entertainment-industry-sectors/, 2024.

[101] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2), Mar. 2018. ISSN 1073-0516. doi: 10.1145/3185593. URL https://doi.org/10.1145/3185593.

[102] S. Jhaver, A. Bruckman, and E. Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi: 10.1145/3359252. URL https://doi.org/10.1145/3359252.

[103] T. D. Jorgensen, S. Pornprasertmanit, A. M. Schoemann, and Y. Rosseel. *semTools: Useful tools for structural equation modeling*, 2026. URL https://CRAN.R-project.org/package=semTools. R package version 0.5-8.

[104] D. Jovic, R. Galwey, L. A. Sparrow, M.-A. Butt, Y. Song, S. Wang-Wills, and E. A. Oliveira. The ai ally project: Designing a human-centred ai tool for women's safety in online spaces. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2791–2795, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3716878. URL https://doi.org/10.1145/3701716.3716878.

[105] L. K. Kaye, A. Orben, D. A. Ellis, S. C. Hunter, and S. Houghton. The conceptual and methodological mayhem of "screen time". *International Journal of Environmental Research and Public Health*, 17(10), 2020. ISSN 1660-4601. doi: 10.3390/ijerph17103661. URL https://www.mdpi.com/1660-4601/17/10/3661.

[106] J. D. Karch. lavaangui: A web-based graphical interface for specifying lavaan models by drawing path diagrams. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–12, 2025. doi: 10.1080/10705511.2024.2420678.

[107] L. S. Kaye, L.-a. M. Hellsten, L. J. McIntyre, and B. P. Hendry. 'there's a fine line between trash-talking and cyberbullying': a qualitative exploration of youth perspectives of online gaming culture. *International Review of Sociology*, 32(3):426 – 442, 2022. doi: 10.1080/03906701.2022.2133407. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141017900&doi=10.1080%2f03906701.2022.2133407&partnerID=40&md5=6d0d6f158f9b1b121788ae8a283b3fe3. Cited by: 1.

[108] P. C. G. Kimberly Magoon, Mary-Jo Robinson, Alexandra Kissling, Victory Ozeua. Best practice for demographic data collection & reporting: Evaluator's guide. Web page, 2024. URL https://publicconsultinggroup.com/wp-content/uploads/2023/10/demographic-data-collection-and-reporting_brief.pdf.

[109] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018. doi: 10.1257/pandp.20181018. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20181018.

[110] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 04 2019. ISSN 2161-7201. doi: 10.1093/jla/laz001. URL https://doi.org/10.1093/jla/laz001.

[111] C. Klemm and W. Pieters. Game mechanics and technological mediation: an ethical perspective on the effects of mmorpg's. *Ethics and Information Technology*, 19

(2):81–93, 2017. doi: 10.1007/s10676-017-9416-6. URL https://doi.org/10.1007/s10676-017-9416-6.

[112] T. Klimstra. Adolescent personality development and identity formation. *Child Development Perspectives*, 7(2):80–84, 2013.

[113] H. Koh, D. Kim, M. Lee, and K. Jung. Can llms recognize toxicity? a structured investigation framework and toxicity metric, 2024. URL https://arxiv.org/abs/2402.06900.

[114] G. Komaç and K. Çağıltay. An overview of trolling behavior in online spaces and gaming context. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–4, 2019. doi: 10.1109/UBMYK48245.2019.8965625.

[115] B. Kordyaka. Digital poison " approaching a theory of toxic behavior in moba games. In *International Conference on Interaction Sciences*, 2018.

[116] B. Kordyaka. Designed for hostility? toxic behavior and ambivalent game elements in online multiplayer game genres. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791085. URL https://doi.org/10.1145/3772318.3791085.

[117] B. Kordyaka and B. Kruse. Curing toxicity – developing design principles to buffer toxic behaviour in massive multiplayer online games. *Safer Communities*, 20(3):133–149, 2021. URL https://www.proquest.com/scholarly-journals/curing-toxicity-developing-design-principles/docview/2579482809/se-2. Copyright - © Emerald Publishing Limited 2021; Last updated - 2023-01-05.

[118] B. Kordyaka, M. Klesel, and K. Jahn. Perpetrators in league of legends: Scale development and validation of toxic behavior. volume 2019-January, page 2486 – 2495, 2019. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083366221&partnerID=40&md5=c3dbc80be58d20cde96d07e8b6c2fb6a. Cited by: 9.

[119] B. Kordyaka, K. Jahn, and B. Niehaves. Towards a unified theory of toxic behavior in video games. *Internet Research*, 30(4):1081–1102, 2020. doi: 10.1108/INTR-08-2019-0343. URL https://doi.org/10.1108/INTR-08-2019-0343.

[120] B. Kordyaka, S. Laato, K. Jahn, J. Hamari, and B. Niehaves. The cycle of toxicity: Exploring relationships between personality and player roles in toxic behavior in multiplayer online battle arena games. *Proc. ACM Hum.-Comput. Interact.*, 7(CHI PLAY), oct 2023. doi: 10.1145/3611043. URL https://doi.org/10.1145/3611043.

[121] B. Kordyaka, S. Laato, A. Stelter, and B. Niehaves. Dark desires? using the theory of basic desires to better understand toxic behavior in multiplayer online games. volume 2023-January, page 5551 – 5559, 2023. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152125652&partnerID=40&md5=5b04a9ef99c3797e74652cb7712e7472. Cited by: 0.

[122] B. Kordyaka, S. Laato, S. Weber, and B. Niehaves. What constitutes victims of toxicity - identifying drivers of toxic victimhood in multiplayer online battle arena games. *Frontiers in Psychology*, 14, 2023. doi: 10.3389/fpsyg.2023.1193172. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164386244&doi=10.3389%2ffpsyg.2023.1193172&partnerID=40&md5=72c2874eefe48d1eb7b2021ebfc8f504. Cited by: 0; All Open Access, Gold Open Access.

[123] B. Kordyaka, S. Park, J. Krath, and S. Laato. Exploring the relationship between offline cultural environments and toxic behavior tendencies in multiplayer online games. *Trans. Soc. Comput.*, 6(1–2), jun 2023. doi: 10.1145/3580346. URL https://doi.org/10.1145/3580346.

[124] Y. Kou. Toxic behaviors in team-based competitive gaming: The case of league of legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, pages 81–92, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380744. doi: 10.1145/3410404.3414243. URL https://doi.org/10.1145/3410404.3414243.

[125] Y. Kou. Punishment and its discontents: An analysis of permanent ban in an online game community. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3476075. URL https://doi.org/10.1145/3476075.

[126] Y. Kou and X. Gui. Playing with strangers: Understanding temporary teams in league of legends. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '14, pages 161–169, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330145. doi: 10.1145/2658537.2658538. URL https://doi.org/10.1145/2658537.2658538.

[127] Y. Kou and X. Gui. When code governs community. In T. Bui and R. Sprague, editors, *PROCEEDINGS OF THE 50TH ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES*, pages 2056–2064, 2017. 50th Annual Hawaii International Conference on System Sciences(HICSS), HI, JAN 03-07, 2017.

[128] Y. Kou and X. Gui. Mediating community-ai interaction through situated explanation: The case of ai-led moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415173. URL https://doi.org/10.1145/3415173.

[129] Y. Kou and X. Gui. Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445279. URL https://doi.org/10.1145/3411764.3445279.

[130] Y. Kou and B. A. Nardi. Governance in league of legends: A hybrid system. In *International Conference on Foundations of Digital Games*, 2014.

[131] Y. Kou, X. Gui, and Y. M. Kow. Ranking practices and distinction in league of legends. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in*

Play, CHI PLAY '16, pages 4–9, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344562. doi: 10.1145/2967934.2968078. URL https://doi.org/10.1145/2967934.2968078.

[132] Y. Kou, X. Gui, S. Zhang, and B. Nardi. Managing disruptive behavior through non-hierarchical governance: Crowdsourcing in league of legends and weibo. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), dec 2017. doi: 10.1145/3134697. URL https://doi.org/10.1145/3134697.

[133] Y. Kou, M. Johansson, and H. Verhagen. Prosocial behavior in an online game community: An ethnographic study. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450353199. doi: 10.1145/3102071.3102078. URL https://doi.org/10.1145/3102071.3102078.

[134] Y. Kou, Y. Li, X. Gui, and E. Suzuki-Gill. 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*, CHI '18, pages 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174152. URL https://doi.org/10.1145/3173574.3174152.

[135] Y. Kou, R. H. L. Hernandez, and X. Gui. "the system is made to inherently push child gambling in my opinion": Child safety, monetization, and moderation on roblox. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713170. URL https://doi.org/10.1145/3706598.3713170.

[136] R. Kowert. Dark participation in games. *Front Psychol*, 11:598947, 2020. ISSN 1664-1078 (Print); 1664-1078 (Electronic); 1664-1078 (Linking). doi: 10.3389/fpsyg.2020.598947.

[137] R. Kowert and C. L. Cook. The toxicity of our (sim) cities: Prevalence of dark participation in games and perceived effectiveness of reporting tools. volume 2022-January, page 3180 – 3189, 2022. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141136398&partnerID=40&md5=a5c1047629ed9671f6914ff959d41c6c. Cited by: 2.

[138] R. Kowert, A. Martel, and W. B. Swann. Not just a game: Identity fusion and extremism in gaming cultures. *Frontiers in Communication*, 7:1007128, 2022.

[139] J. L. Kröger, P. Raschke, J. Percy Campbell, and S. Ullrich. Surveilling the gamers: Privacy impacts of the video game industry. *Entertainment Computing*, 44:100537, 2023. ISSN 1875-9521. doi: https://doi.org/10.1016/j.entcom.2022.100537. URL https://www.sciencedirect.com/science/article/pii/S187595212200060X.

[140] J. H. Kuznekoff and L. M. Rose. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society*, 15(4):541–556, 2013. doi: 10.1177/1461444812458271. URL https://doi.org/10.1177/1461444812458271.

[141] H. Kwak, J. Blackburn, and S. Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3739–3748, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702529. URL https://doi.org/10.1145/2702123.2702529.

[142] J. Ladanyi and S. Doyle-Portillo. The development and validation of the grief play scale (gps) in mmorpgs. *PERSONALITY AND INDIVIDUAL DIFFERENCES*, 114:125–133, AUG 1 2017. doi: 10.1016/j.paid.2017.03.062.

[143] P. Lankoski and S. Björk. Gameplay design patterns for believable non-player characters. In *Proceedings of DiGRA 2007 Conference: Situated Play*, Tampere, 2007. DiGRA.

[144] B. Latour. 10 "where are the missing masses? the sociology of a few mundane artifacts". 1992.

[145] R. S. Lazarus and S. Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[146] R. S. Lazarus and S. Folkman. Transactional theory and research on emotions and coping. *European Journal of Personality*, 1(3):141–169, 1987. doi: 10.1002/per.2410010304. URL https://doi.org/10.1002/per.2410010304.

[147] M. Lea and R. Spears. Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies*, 34(2):283–301, 1991. ISSN 0020-7373. doi: https://doi.org/10.1016/0020-7373(91)90045-9. URL https://www.sciencedirect.com/science/article/pii/0020737391900459. Special Issue: Computer-supported Cooperative Work and Groupware. Part 1.

[148] A.-D. League. Hate is no game: Harassment and positive social experiences in online games 2022, 2022. URL https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022.

[149] M. Lee, D. Johnson, P. Tanjitpiyanond, and W. R. Louis. It's habit, not toxicity, driving hours spent in dota 2. *ENTERTAINMENT COMPUTING*, 41, MAR 2022. ISSN 1875-9521. doi: 10.1016/j.entcom.2021.100472.

[150] M. K. Lee, A. Jain, H. J. Cha, S. Ojha, and D. Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi: 10.1145/3359284. URL https://doi.org/10.1145/3359284.

[151] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi: 10.1145/3359283. URL https://doi.org/10.1145/3359283.

[152] O. Leis, F. Lautenbach, P. D. Birch, and A.-M. Elbe. Stressors, associated responses, and coping strategies in professional esports players: A qualitative study. *International Journal of Esports*, 1, 2022. URL https://www.ijesports.org/article/76/html.

[153] M. Lemercier-Dugarin, L. Romo, C. Tijus, and O. Zerhouni. "who are the cyka blyat?" how empathy, impulsivity, and motivations to play predict aggressive behaviors in multiplayer online games. *CYBERPSYCHOLOGY BEHAVIOR AND SOCIAL NETWORKING*, 24(1):63–69, JAN 1 2021. ISSN 2152-2715. doi: 10.1089/cyber.2020.0041.

[154] D. J. Leonard. "live in your world, play in ours": Race, video games, and consuming the other. *Simile: Studies in Media & Information Literacy Education*, 3:1–9, 2003.

[155] D. J. Leonard. Not a hater, just keepin' it real: The importance of race- and gender-based game studies. *Games and Culture*, 1(1):83–88, 2023/05/09 2006. doi: 10.1177/1555412005281910. URL https://doi.org/10.1177/1555412005281910.

[156] I. J. Livingston, C. Gutwin, R. L. Mandryk, and M. Birk. How players value their characters in world of warcraft. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 1333–1343, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325400. doi: 10.1145/2531602.2531661. URL https://doi.org/10.1145/2531602.2531661.

[157] S. Livingstone and E. J. Helsper. Parental mediation of children's internet use. *Journal of Broadcasting & Electronic Media*, 52(4):581–599, 2008. doi: 10.1080/08838150802437396. URL https://doi.org/10.1080/08838150802437396.

[158] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson. Risks and safety on the internet: the perspective of european children: full findings and policy implications from the eu kids online survey of 9-16 year olds and their parents in 25 countries. Technical report, London, UK, 2011. URL http://www.eukidsonline.net/.

[159] R. Ma, Y. Li, and Y. Kou. Transparency, fairness, and coping: How players experience moderation in multiplayer online games. In *PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI 2023*, 2023. doi: 10.1145/3544548.3581097. CHI conference on Human Factors in Computing Systems (CHI), Hamburg, GERMANY, APR 23-28, 2023.

[160] R. Ma, Y. You, X. Gui, and Y. Kou. How do users experience moderation?: A systematic literature review. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), Oct. 2023. doi: 10.1145/3610069. URL https://doi.org/10.1145/3610069.

[161] D. MacKenzie and J. Wajcman. *The Social Shaping of Technology*. Oxford University Press, United Kingdom, 1999. ISBN 0-335-19914-3.

[162] R. Maghsoudi, J. Shapka, and P. Wisniewski. Examining how online risk exposure and online social capital influence adolescent psychological stress. *Computers in Human Behavior*, 113:106488, 2020. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2020.106488. URL https://www.sciencedirect.com/science/article/pii/S0747563220302405.

[163] B. Maher. Can a video game company tame toxic behaviour? *Nature*, 531(7596): 568–571, Mar 2016. ISSN 1476-4687 (Electronic); 0028-0836 (Linking). doi: 10.1038/531568a.

[164] E. Martellozzo. *Online Child Sexual Abuse: Grooming, Policing and Child Protection in a Multi-Media World - Endorsements*. Routledge, USA, 1st edition, 2012. ISBN 0415618215.

[165] M. Märtens, S. Shen, A. Iosup, and F. Kuipers. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6, 2015. ISBN 2156-8146. doi: 10.1109/NetGames.2015.7382991.

[166] L. McLean and M. D. Griffiths. Female gamers' experience of online harassment and social support in online gaming: A qualitative study. *International Journal of Mental Health and Addiction*, 17(4):970–994, 2019. doi: 10.1007/s11469-018-9962-0. URL https://doi.org/10.1007/s11469-018-9962-0.

[167] R. . I. McMaster University. Guidance on demographic questions. Web page, 2025. URL https://research.mcmaster.ca/home/support-for-researchers/ethics/mcmaster-research-ethics-board-mreb/guidance-on-demographic-questions/. McMaster Research Ethics Board (MREB).

[168] D. Melhart, J. Togelius, B. Mikkelsen, C. Holmgård, and G. N. Yannakakis. The ethics of ai in games. *IEEE Transactions on Affective Computing*, 15(1):79–92, 2024. doi: 10.1109/TAFFC.2023.3276425.

[169] O. J. Merry and K. C. Whitfield. The landscape of sexual harm in the video game, streaming, and esports community. *Journal of Interpersonal Violence*, 40(13-14):3307–3332, 2025. doi: 10.1177/08862605241271349. URL https://doi.org/10.1177/08862605241271349. PMID: 39169727.

[170] G. S. Mesch. Parental mediation, online activities, and cyberbullying. *CyberPsychology & Behavior*, 12(4):387–393, 2009. doi: 10.1089/cpb.2009.0068. URL https://journals.sagepub.com/doi/abs/10.1089/cpb.2009.0068.

[171] D. L. Morgan. Focus groups. *Annual Review of Sociology*, 22(Volume 22, 1996):129–152, 1996. ISSN 1545-2115. doi: https://doi.org/10.1146/annurev.soc.22.1.129. URL https://www.annualreviews.org/content/journals/10.1146/annurev.soc.22.1.129.

[172] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*, 76:197–213, 2018. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2018.02.016. URL https://www.sciencedirect.com/science/article/pii/S0167404818301597.

[173] L. Nakamura. *"It's a Nigger in Here! Kill the Nigger!"*. 2023/05/09 2012. ISBN 9781444361506. doi: https://doi.org/10.1002/9781444361506.wbiems159. URL https://doi.org/10.1002/9781444361506.wbiems159.

[174] B. Nardi and J. Harris. Strangers and friends: collaborative play in world of warcraft. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, page 149–158, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932496. doi: 10.1145/1180875.1180898. URL https://doi.org/10.1145/1180875.1180898.

[175] J. A. M. Neto, K. M. Yokoyama, and K. Becker. Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 26–33, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349512. doi: 10.1145/3106426.3106452. URL https://doi.org/10.1145/3106426.3106452.

[176] F. Neumayer, V. Jantzer, S. Lerch, F. Resch, and M. Kaess. Traditional bullying and cyberbullying victimization independently predict changes in problematic internet gaming in a longitudinal sample. *Journal of Adolescent Health*, 73(2):288 – 295, 2023. doi: 10.1016/j.jadohealth.2023.03.013. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163125521&doi=10.1016%2fj.jadohealth.2023.03.013&partnerID=40&md5=c555c1a4e49ae2bf318079707479a225. Cited by: 0.

[177] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883062. URL https://doi.org/10.1145/2872427.2883062.

[178] L. M. Padilla-Walker, S. M. Coyne, A. M. Fraser, W. J. Dyer, and J. B. Yorgason. Parents and adolescents growing up in the digital age: Latent growth curve analysis of proactive media monitoring. *Journal of Adolescence*, 35(5):1153–1165, 2012. doi: https://doi.org/10.1016/j.adolescence.2012.03.005. URL https://onlinelibrary.wiley.com/doi/abs/10.1016/j.adolescence.2012.03.005.

[179] M. Pappas. Modulate and rockstar games partner to enhance player experience in gta online. Web page, 2024. URL https://www.modulate.ai/blog/rockstar-gta-online.

[180] J. Park, I. Lediaeva, M. Lopez, A. Godfrey, K. C. Madathil, H. Zinzow, and P. Wisniewski. How affordances and social norms shape the discussion of harmful social media

challenges on reddit. *Human Factors in Healthcare*, 3:100042, 2023. ISSN 2772-5014. doi: https://doi.org/10.1016/j.hfh.2023.100042. URL https://www.sciencedirect.com/science/article/pii/S277250142300009X.

[181] A. K. Pasayat, M. Shrestha, and P. P. Priya. Unleashing the possibilities of play: Analyzing the positive and negative consequences of multiplayer online gaming. *Entertainment Computing*, 52:100898, 2025. ISSN 1875-9521. doi: https://doi.org/10.1016/j.entcom.2024.100898. URL https://www.sciencedirect.com/science/article/pii/S1875952124002660.

[182] C. J. Passmore and R. L. Mandryk. A taxonomy of coping strategies and discriminatory stressors in digital gaming. *Frontiers in Computer Science*, 2, 2020. ISSN 2624-9898. doi: 10.3389/fcomp.2020.00040. URL https://www.frontiersin.org/articles/10.3389/fcomp.2020.00040.

[183] T. J. Pinch and W. E. Bijker. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14:399–441, 1984. URL http://www.jstor.org/stable/285355.

[184] H. Pink. Snap to character: Building strong player attachment through narrative. In Game Developers Conference., 2017. URL https://www.gdcvault.com/play/1024625/Snap-to-Character-Building-Strong.

[185] A. T. Pinter, P. J. Wisniewski, H. Xu, M. B. Rosson, and J. M. Caroll. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '17, page 352–357, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349215. doi: 10.1145/3078072.3079722. URL https://doi.org/10.1145/3078072.3079722.

[186] F. M. Plaza-del arco, D. Nozza, and D. Hovy. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Y.-L. Chung, P. Röttger, D. Nozza, Z. Talat, and A. Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.6. URL https://aclanthology.org/2023.woah-1.6/.

[187] S. Poeller and R. B. Robinson. Mute, block, punish, reward? a call to shift the research focus from concealing toxicity in games to promoting genuine positive behavior. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY Companion '24, page 282–284, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706929. doi: 10.1145/3665463.3678863. URL https://doi.org/10.1145/3665463.3678863.

[188] D. Poulus, T. J. Coulter, M. G. Trotter, and R. Polman. Stress and coping in esports and the influence of mental toughness. *Frontiers in Psychology*, Volume 11 - 2020, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.00628. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.00628.

[189] L. Qing. When cyberbullying and bullying meet gaming: A systemic review of the literature. *Journal of Psychology & Psychotherapy*, 5:1–11, 2015. doi: 10.4172/2161-0487.1000195. URL https://www.longdom.org/open-access/when-cyberbullying-and-bullying-meet-gaming-a-systemic-review-of-the-literature-21pdf.

[190] Qualtrics Team. *Qualtrics*. Qualtrics, VProvo, Utah, USA, 2005. URL https://www.qualtrics.com.

[191] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[192] A. Razi, K. Badillo-Urquiola, and P. J. Wisniewski. Let's talk about sext: How adolescents seek support and advice about their online sexual experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376400. URL https://doi.org/10.1145/3313831.3376400.

[193] E. Reid, R. L. Mandryk, N. A. Beres, M. Klarkowski, and J. Frommel. Feeling good and in control: In-game tools to support targets of toxicity. *Proc. ACM Hum.-Comput. Interact.*, 6(CHI PLAY), oct 2022. doi: 10.1145/3549498. URL https://doi.org/10.1145/3549498.

[194] A. Romhanyi, G. Wells, and C. Steinkuehler. Laugh, mute, or report? how young adults handle toxicity in multiplayer games, 2025. Manuscript under review in ACM Games: Research and Practice.

[195] Y. Rosseel, T. D. Jorgensen, and L. De Wilde. *lavaan: Latent Variable Analysis*, 2025. URL https://CRAN.R-project.org/package=lavaan. R package version 0.6-20.

[196] S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24, 2020. doi: 10.1109/TAFFC.2017.2761757.

[197] K. Salen and E. Zimmerman. *Rules of Play: Game Design Fundamentals*. MIT Press, Cambridge, Mass., 2003. ISBN 9780262240451.

[198] C. Salmivalli. Participant role approach to school bullying: implications for interventions. *Journal of Adolescence*, 22(4):453–459, 1999. ISSN 0140-1971. doi: https://doi.org/10.1006/jado.1999.0239. URL https://www.sciencedirect.com/science/article/pii/S0140197199902397.

[199] A. Salter and B. Blodgett. Toxic geek masculinity in media. *Toxic Geek Masculinity in Media: Sexism, Trolling, and Identity Policing*, 11 2017. doi: 10.1007/978-3-319-66077-6.

[200] L. Schlegel and R. Kowert. *Gaming and extremism: The radicalization of digital playgrounds*. Taylor & Francis, 2024.

[201] S. K. Schneider, L. O'donnell, A. Stueve, and R. W. S. Coulter. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American Journal of Public Health*, 102(1):171 – 177, 2012. doi: 10.2105/AJPH.2011.300308. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84857311826&doi=10.2105%2fAJPH.2011.300308&partnerID=40&md5=52d156922ab42dcee6b86ea2edfe46b6. Cited by: 612; All Open Access, Green Open Access.

[202] K. Schulenberg, L. Li, G. Freeman, S. Zamanifard, and N. J. McNeese. 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*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581090. URL https://doi.org/10.1145/3544548.3581090.

[203] S. Şengün, J. Salminen, P. Mawhorter, S.-g. Jung, and B. Jansen. Exploring the relationship between game content and culture-based toxicity: A case study of league of legends and mena players. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 87–95, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368858. doi: 10.1145/3342220.3343652. URL https://doi.org/10.1145/3342220.3343652.

[204] S. Şengün, J. M. Santos, J. Salminen, S. gyo Jung, and B. J. Jansen. Do players communicate differently depending on the champion played? exploring the proteus effect in league of legends. *Technological Forecasting and Social Change*, 177:121556, 2022. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2022.121556. URL https://www.sciencedirect.com/science/article/pii/S0040162522000889.

[205] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.244. URL https://aclanthology.org/2023.acl-long.244/.

[206] C. Shen, Q. Sun, T. Kim, G. Wolff, R. Ratan, and D. Williams. Viral vitriol: Predictors and contagion of online toxicity in world of tanks. *Computers in Human Behavior*, 108:106343, 2020. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2020.106343. URL https://www.sciencedirect.com/science/article/pii/S0747563220300972.

[207] J. Sherman and A. P. McClard. Toxicity v. toxicity: How ethnography can inform scalable technical solutions. volume 2020, pages 279–297, 10 2020. doi: 10.1111/epic.12037.

[208] A. Shoshani and M. Krauskopf. The fortnite social paradox: The effects of violent-cooperative multi-player video games on children's basic psychological needs and prosocial behavior. *Computers in Human Behavior*, 116:106641, 2021. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2020.106641. URL https://www.sciencedirect.com/science/article/pii/S0747563220303885.

[209] V. J. Shute, M. Ventura, and F. Ke. The power of play: The effects of portal 2 and lumosity on cognitive and noncognitive skills. *Computers & Education*, 80:58–67, 2015. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2014.08.013. URL https://www.sciencedirect.com/science/article/pii/S0360131514001869.

[210] V. Slavtcheva-Petkova, V. J. Nash, and M. Bulger. Evidence on the extent of harms experienced by children as a result of online risks: implications for policy and research. *Information, Communication & Society*, 18(1):48–62, 2015. doi: 10.1080/1369118X.2014.934387. URL https://doi.org/10.1080/1369118X.2014.934387.

[211] P. Slovak, K. Salen, S. Ta, and G. Fitzpatrick. Mediating conflicts in minecraft: Empowering learning in online multiplayer games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174169. URL https://doi.org/10.1145/3173574.3174169.

[212] M. Smith, P. Birch, and D. Bright. Identifying stressors and coping strategies of elite esports competitors. *International Journal of Gaming and Computer-Mediated Simulations*, 11:22–39, 04 2019. doi: 10.4018/IJGCMS.2019040102.

[213] A. Souza, J. Pegorini, N. Yada, L. Costa, and F. C. Souza. Exploring toxic behavior in multiplayer online games: perceptions of different genders. 10 2021.

[214] L. Sparrow, M. Gibbs, and M. Arnold. Apathetic villagers and the trolls who love them: Player amorality in online multiplayer games. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, OzCHI '19, pages 447–451, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450376969. doi: 10.1145/3369457.3369514. URL https://doi.org/10.1145/3369457.3369514.

[215] L. Sparrow, M. Gibbs, and M. Arnold. The ethics of multiplayer game design and community management: Industry perspectives and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445363. URL https://doi.org/10.1145/3411764.3445363.

[216] L. A. Sparrow, R. Galwey, D. Jovic, T. Hardwick, and M.-A. Butt. Towards ethical ai moderation in multiplayer games. *Proc. ACM Hum.-Comput. Interact.*, 8(CHI PLAY), Oct. 2024. doi: 10.1145/3677109. URL https://doi.org/10.1145/3677109.

[217] K. Stapor. *Descriptive and Inferential Statistics*, pages 63–131. Springer International Publishing, Cham, 2020. ISBN 978-3-030-45799-0. doi: 10.1007/978-3-030-45799-0_2. URL https://doi.org/10.1007/978-3-030-45799-0_2.

[218] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445092. URL https://doi.org/10.1145/3411764.3445092.

[219] C. Steinkuehler. Massively multiplayer online gaming as a constellation of literacy practices. *E-learning*, 4, 09 2007. doi: 10.2304/elea.2007.4.3.297.

[220] C. Steinkuehler. Massively multiplayer online games as an educational technology: An outline for research. *Educational Technology*, 48(1):10–21, 2008. ISSN 00131962. URL http://www.jstor.org/stable/44429539.

[221] C. Steinkuehler. Games as social platforms. *ACM Games*, 1(1), mar 2023. doi: 10.1145/3582930. URL https://doi.org/10.1145/3582930.

[222] C. Steinkuehler and S. Duncan. Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology*, 17:530–543, 12 2008. doi: 10.1007/s10956-008-9120-8.

[223] C. A. Steinkuehler and D. Williams. Where everybody knows your (screen) name: Online games as "third places". *Journal of computer-mediated communication*, 11(4): 885–909, 2006.

[224] W. Stoop, F. Kunneman, A. van den Bosch, and B. Miller. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24, Florence, Italy, aug 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3503. URL https://aclanthology.org/W19-3503.

[225] J. Suler. The online disinhibition effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7:321–6, 07 2004. doi: 10.1089/1094931041291295.

[226] W. Y. Tang and J. Fox. Men's harassment behavior in online video games: Personality traits and game factors. *Aggressive Behavior*, 42(6):513–521, 2016. doi: https://doi.org/10.1002/ab.21646. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21646.

[227] W. Y. Tang, F. Reer, and T. Quandt. Investigating sexual harassment in online video games: How personality and context factors are related to toxic sexual behaviors against fellow players. *Aggressive Behavior*, 46(1):127–135, 2020. doi: https://doi.org/10.1002/ab.21873. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21873.

[228] T. L. Taylor. Beyond management: Considering participatory design and governance in player culture. *First Monday*, 7(0), 2023/04/30 2006. doi: 10.5210/fm.v0i0.1611. URL https://firstmonday.org/ojs/index.php/fm/article/view/1611.

[229] K. S. Tekinbaş, K. Jagannath, U. Lyngs, and P. Slovák. Designing for youth-centered moderation and community governance in minecraft. *ACM Trans. Comput.-Hum. Interact.*, 28(4), July 2021. ISSN 1073-0516. doi: 10.1145/3450290. URL https://doi.org/10.1145/3450290.

[230] R. H. Thaler, C. R. Sunstein, and J. P. Balz. Choice architecture. April 2010. URL https://ssrn.com/abstract=2536504orhttp://dx.doi.org/10.2139/ssrn.2536504.

[231] J. J. Thompson, B. H. Leung, M. R. Blair, and M. Taboada. Sentiment analysis of player chat messaging in the video game starcraft 2: Extending a lexicon-based model. *Knowledge-Based Systems*, 137:149–162, 2017. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.09.022. URL https://www.sciencedirect.com/science/article/pii/S0950705117304240.

[232] R. G. TimTamMonster. /dev: Making behavioral systems matter more. Web page, 2022. URL https://www.leagueoflegends.com/en-us/news/dev/dev-making-behavioral-systems-matter-more/.

[233] S. Tomkinson and B. van den Ende. 'thank you for your compliance': Overwatch as a disciplinary system. *Games and Culture*, 17(2):198–218, 2023/04/30 2021. doi: 10.1177/15554120211026257. URL https://doi.org/10.1177/15554120211026257.

[234] S. Türkay and S. Adinolf. Friending to flame: How social features affect player behaviours in an online collectible card game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300567. URL https://doi.org/10.1145/3290605.3300567.

[235] S. Türkay, J. Formosa, S. Adinolf, R. Cuthbert, and R. Altizer. See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376191. URL https://doi.org/10.1145/3313831.3376191.

[236] A. Tyack, P. Wyeth, and D. Johnson. The appeal of moba games: What makes people start, stay, and stop. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, pages 313–325, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344562. doi: 10.1145/2967934.2968098. URL https://doi.org/10.1145/2967934.2968098.

[237] P.-P. Verbeek. Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3):361–380, 2023/05/06 2006. doi: 10.1177/0162243905285847. URL https://doi.org/10.1177/0162243905285847.

[238] S. Vie. *Policies, Terms of Service, and Social Networking Games*, pages 54–68. Routledge, 10 2015. ISBN 9781315748825. doi: 10.4324/9781315748825-5.

[239] E. A. Vogels. Teens and cyberbullying 2022. Pew Research Center, 2022. URL https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/.

[240] S. Wachs, M. F. Wright, and A. T. Vazsonyi. Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behaviour and Mental Health*, 29(3):179–188, 2019. doi: https://doi.org/10.1002/cbm.2116. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cbm.2116.

[241] G. Wang, J. Zhao, M. Van Kleek, and N. Shadbolt. Protection or punishment? relating the design space of parental control apps and perceptions about them to support parenting for online safety. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), Oct. 2021. doi: 10.1145/3476084. URL https://doi.org/10.1145/3476084.

[242] G. Wells, A. Romhanyi, J. G. Reitman, R. Gardner, K. Squire, and C. Steinkuehler. Right-wing extremism in mainstream games: A review of the literature. *Games and Culture*, 19(4):469–492, 2023. doi: 10.1177/15554120231167214. URL https://doi.org/10.1177/15554120231167214.

[243] G. Wells, A. Romhanyi, and C. Steinkuehler. Hate speech and hate-based harassment in online games. *Frontiers in Psychology*, Volume 15 - 2024, 2025. ISSN 1664-1078. doi: 10.3389/fpsyg.2024.1422422. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1422422.

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

[245] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, and P. Software. *dplyr*. URL https://dplyr.tidyverse.org,https://github.com/tidyverse/dplyr. version 1.2.0.

[246] M. Wijkstra, K. Rogers, R. L. Mandryk, R. C. Veltkamp, and J. Frommel. Help, my game is toxic! first insights from a systematic literature review on intervention systems for toxic behaviors in online video games. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY Companion '23, page 3–9, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700293. doi: 10.1145/3573382.3616068. URL https://doi.org/10.1145/3573382.3616068.

[247] M. Wijkstra, K. Rogers, R. L. Mandryk, R. C. Veltkamp, and J. Frommel. How to tame a toxic player? a systematic literature review on intervention systems for toxic behaviors in online video games. *Proc. ACM Hum.-Comput. Interact.*, 8(CHI PLAY), Oct. 2024. doi: 10.1145/3677080. URL https://doi.org/10.1145/3677080.

[248] S. Wilkinson. Focus group methodology: a review. *International Journal of Social Research Methodology*, 1(3):181–203, 1998. doi: 10.1080/13645579.1998.10846874. URL https://doi.org/10.1080/13645579.1998.10846874.

[249] L. Winner. Do artifacts have politics? *JSTOR*, 109:121–136, 1980. URL http://www.jstor.org/stable/20024652.

[250] P. Wisniewski, H. Jia, N. Wang, S. Zheng, H. Xu, M. B. Rosson, and J. M. Carroll. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 4029–4038, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702240. URL https://doi.org/10.1145/2702123.2702240.

[251] P. Wisniewski, H. Jia, H. Xu, M. B. Rosson, and J. M. Carroll. "preventative" vs. "reactive": How parental mediation influences teens' social media privacy behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 302–316, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450329224. doi: 10.1145/2675133.2675293. URL https://doi.org/10.1145/2675133.2675293.

[252] P. Wisniewski, H. Xu, M. B. Rosson, D. F. Perkins, and J. M. Carroll. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3919–3930, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858317. URL https://doi.org/10.1145/2858036.2858317.

[253] P. Wisniewski, A. K. Ghosh, H. Xu, M. B. Rosson, and J. M. Carroll. Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety? In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 51–69, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi: 10.1145/2998181.2998352. URL https://doi.org/10.1145/2998181.2998352.

[254] S. Wong and R. Ratan. How does exposure to general and sexual harassment relate to female gamers' coping strategies and mental health? *GAMES AND CULTURE*, 2023 AUG 1 2023. doi: 10.1177/15554120231177600.

[255] M. Wu. *Examining Tilt as Emotional Dysregulation in Gaming*. PhD thesis, UC Irvine, 2025.

[256] M. Wu, J. S. Lee, and C. Steinkuehler. Understanding tilt in esports: A study on young league of legends players. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445143. URL https://doi.org/10.1145/3411764.3445143.

[257] S. Xiao, S. Jhaver, and N. Salehi. Addressing harm in online gaming communities – the opportunities and challenges for a restorative justice approach, 2022. URL https://doi.org/10.48550/arXiv.2211.01524.

[258] S. Xiao, S. Jhaver, and N. Salehi. Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach. *ACM Trans. Comput.-Hum. Interact.*, 30(6), sep 2023. ISSN 1073-0516. doi: 10.1145/3603625. URL https://doi.org/10.1145/3603625.

[259] Z. Yang, N. Grenon-Godbout, and R. Rabbany. Game on, hate off: A study of toxicity in online multiplayer environments. *ACM Games*, 2(2), Aug. 2024. doi: 10.1145/3675805. URL https://doi.org/10.1145/3675805.

[260] F. Yang-Wallentin, K. G. Jöreskog, and H. Luo. Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3):392–423, 2010. doi: 10.1080/10705511.2010.489003. URL https://www.tandfonline.com/doi/abs/10.1080/10705511.2010.489003.

[261] M. L. Ybarra, D. Finkelhor, K. J. Mitchell, and J. Wolak. Associations between blocking, monitoring, and filtering software on the home computer and youth-reported unwanted exposure to sexual material online. *Child Abuse & Neglect*, 33(12):857–869, 2009. ISSN 0145-2134. doi: https://doi.org/10.1016/j.chiabu.2008.09.015. URL https://www.sciencedirect.com/science/article/pii/S014521340900218X.

[262] A. Z. H. Yee and B. Luyt. Who benefits from the idea of screen time? *New Media & Society*, 0(0):14614448261444717, 2026. doi: 10.1177/14614448261444717. URL https://doi.org/10.1177/14614448261444717.

[263] B. Yim, A. Lepp, B. Dowdell, and J. E. Barkley. The gamer identity scale: A measure of self concept as a video gamer. *Computers in Human Behavior*, 138:107476, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2022.107476. URL https://www.sciencedirect.com/science/article/pii/S0747563222002965.

[264] H. Yu, Z. Yang, K. Pelrine, J. Godbout, and R. Rabbany. Open, closed, or small language models for text classification?, 08 2023.

[265] Z.-J. Zhong. The effects of collective mmorpg (massively multiplayer online role-playing games) play on gamers' online and offline social capital. *Comput. Hum. Behav.*, 27(6):2352–2363, Nov. 2011. ISSN 0747-5632. doi: 10.1016/j.chb.2011.07.014. URL https://doi.org/10.1016/j.chb.2011.07.014.

[266] Z. Zhu, R. Zhang, and Y. Qin. Toxicity and prosocial behaviors in massively multiplayer online games: The role of mutual dependence, power, and passion. *Journal of Computer-Mediated Communication*, 27(6), 09 2022. ISSN 1083-6101. doi: 10.1093/jcmc/zmac017. URL https://doi.org/10.1093/jcmc/zmac017.

[267] T. Zioga, C. Nega, P. Roussos, and P. Kourtesis. Validation of the gaming skills questionnaire in adolescence: Effects of gaming skills on cognitive and affective

functioning. *European Journal of Investigation in Health, Psychology and Education*, 14(3):722–752, 2024. ISSN 2254-9625. doi: 10.3390/ejihpe14030048. URL https://www.mdpi.com/2254-9625/14/3/48.

[268] Á. Zsila, R. Shabahang, M. S. Aruguete, and G. Orosz. Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences. *Aggressive Behavior*, 48(3):356–364, 2022. doi: https://doi.org/10.1002/ab.22023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.22023.

[269] S. Şengün, J. Soongyo, B. J. Jansen, J. Salminen, and P. Mawhorter. Analyzing hate speech toward players from the mena in league of legends. 2019. doi: 10.1145/3290607.3312924. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067311747&doi=10.1145%2f3290607.3312924&partnerID=40&md5=4f539c11ac99ac9cabc75c435dd9d822. Cited by: 3.

Appendices

Appendix

Young Adults' Moderation Tools Perception

A Survey Instrument

A.1 Demographics

1. Age (in years) (Number)
2. Gender (Single select)
    - Male
    - Female
    - Nonbinary
    - Prefer to self-describe [ ]
    - Prefer not to answer

3. Level of education (Single select)
    - High School
    - Some college
    - Bachelor's degree
    - Some graduate classes
    - Graduate Degree
    - GED
    - Prefer not to answer
4. Which of the following best describes your sexual orientation? (Single select)
    - Heterosexual (straight)
    - Homosexual (gay)
    - Bisexual
    - Asexual
    - Prefer to self-describe [ ]
    - Prefer not to answer
5. Ethnicity (Single select)
    - Asian
    - Black/African

- White/Caucasian
- Hispanic/Latinx
- Native American
- Pacific Islander
- Prefer to self-describe [ ]
- Prefer not to answer

A.2 Game Experience

- What video game genres do you usually play? Select all that apply.
  - Sports/racing (ex. FIFA, NBA2K, Forza)
  - Multiplayer online battle arena (MOBA ex. League of Legends, DOTA 2)
  - Shooter (ex. Call of Duty, Overwatch, CS:GO, Valorant, Fortnite PUBG)
  - Fighting (ex. Street Fighter, Tekken, Super Smash Bros.)
  - Massive multiplayer online (MMO ex. World of Warcraft, FF14, ESO, GTA)
  - Real-time strategy (RTS ex. Starcraft, Age of Empires, Clash Royale)
  - Turn-based (ex. Civilization)
  - Party Games (ex. Fall Guys, Mario Party, Jackbox)
  - Sandbox (ex. Minecraft, No Man's Sky)
  - Simulation (ex. Sims, Animal Crossing)

- Other:

- How many hours per week do you usually spend playing video games online? (Number)
- For how many years have you been playing video games online? (Number)
- How much do you identify with the term "gamer"? (5-point Likert: Strongly disagree - Strongly agree)
- Please indicate how much you agree with each of the following statements. (5-point Likert: Strongly disagree - Strongly agree)
    1. I am good at playing videogames.
    2. I am skilled at the games I play.
- How often do you play ranked games (games on the competitive ladder)? (5-point Likert: Never - Always)

A.3 Experiences of Toxicity and Responses to It

The following questions are about your experiences with toxicity in online games. By toxicity, we mean negative, harmful, or unpleasant behaviors that can disrupt morale and team dynamics. In gaming, this often includes things like harassing, bullying, cheating, trolling, griefing, and flaming other players.

A.3.1 Bystander Experiences of Toxicity: Intensity & Frequency

Please indicate how often you have encountered the following situations in online games. (5-point Likert: Never - Always)

1. I have witnessed people embarrassing others while playing games online.

2. I have witnessed people using offensive names while playing games online.

3. I have witnessed trolling/griefing [deliberate attempts to upset or provoke another player] while playing games online.

4. I have witnessed harassment while playing games online.

5. I have witnessed people being harassed in an online game based on their membership in a minoritized group.

6. I have witnessed people threatening other players in games online.

7. I know someone who has been doxxed [had their private information published online without consent] by someone they played games online with.

8. I know someone who has been stalked by someone they played games online with.

9. I know someone who has been swatted (had a hoax call to emergency services about them) by someone they played games online with.

A.3.2 Bystander Responses

Please indicate how often you have done the following in online games. (5-point Likert: Never - Always)

1. When I witness toxicity, I take it as a joke and laugh it off.

2. When I witness toxicity, I stop talking or withdraw from chat.

3. When I witness toxicity, I mute or hide the chat and carry on.

4. When I witness toxicity, I refocus the conversation on the game.

5. When I witness toxicity, I reach out to support the victim.

6. When I witness toxicity, I act toxically back to the offender.

7. When I witness toxicity, I leave the match.

8. When I witness toxicity, I ignore it.

9. When I witness toxicity, I report it to the game company.

10. When I witness toxicity, I log off for a while.

11. When I witness toxicity, I stop playing the game completely.

12. When I witness toxicity, I call the perpetrator out for it.

13. When I witness toxicity, I ... (please specify). [other]

A.3.3 Victim Experiences of Toxicity: Intensity & Frequency

Please indicate how often you have encountered the following situations in online games. (5-point Likert: Never - Always)

1. I have been purposefully embarrassed by other people while playing games online.

2. I have had other people call me offensive names while playing games online.

3. I have been trolled/griefed [deliberate attempts to upset or provoke another player] while playing online.

4. I have been the victim of harassment while playing games online.

5. I have been harassed in an online game based on my membership in a minoritized group.

6. I have been threatened by other players while playing games online.

7. I have been doxxed [had my private information published online without my consent] by someone I played games online with.

8. I have been stalked by someone I played games online with.

9. I have been swatted (had a hoax call to emergency services about me) by someone they played games online with.

A.3.4 Victim Responses

Please indicate how often you have done the following in online games. (5-point Likert: Never - Always)

1. When I am the victim of toxicity, I take it as a joke and laugh it off.

2. When I am the victim of toxicity, I stop talking or withdraw from chat.

3. When I am the victim of toxicity, I mute or hide the chat and carry on.

4. When I am the victim of toxicity, I refocus the conversation on the game.

5. When I am the victim of toxicity, I reach out to bystanders for support.

6. When I am the victim of toxicity, I act toxically back to the offender.

7. When I am the victim of toxicity, I leave the match.

8. When I am the victim of toxicity, I ignore it.

9. When I am the victim of toxicity, I report it to the game company.

10. When I am the victim of toxicity, I log off for a while.

11. When I am the victim of toxicity, I stop playing the game completely.

12. When I am the victim of toxicity, I call the perpetrator out for it.

13. When I am the victim of toxicity, I ... (please specify). [other]

A.3.5 Perpetrator Experiences of Toxicity: Intensity & Frequency

Please indicate how often you have done the following in online games. Reminder: No identifying information associated with your survey responses is recorded, so all your answers are completely anonymous. (5-point Likert: Never - Always)

1. I have purposefully embarrassed other people while playing games online.

2. I have called other players offensive names while playing games online.

3. I have trolled/griefed [deliberately attempted to upset or provoke another player] other players while playing games online.

4. I have harassed other players while playing games online.

5. I have harassed other players in an online game based on their membership in a minoritized group.

6. I threatened other players while playing games online.

7. I doxxed [having their private information published online without consent] someone I played games online with.

8. I stalked someone I played games online with.

9. I swatted (made a hoax call to emergency services about them) someone I played games online with.

A.3.6 Perpetrator Responses

Please indicate how often you have encountered the following situations in online games. (5-point Likert: Never - Always)

1. When I get toxic, other players take it as a joke and laugh it off.
2. When I get toxic, other players stop talking or withdraw from chat.
3. When I get toxic, other players mute or hide the chat and carry on.
4. When I get toxic, other players refocus the conversation on the game.
5. When I get toxic, other players support each other.
6. When I get toxic, other players get toxic in return.
7. When I get toxic, other players leave the match.
8. When I get toxic, other players ignore it.
9. When I get toxic, I get reported.
10. When I get toxic, other players log off.
11. When I get toxic, other players call me out for it.
12. When I get toxic, ... (please specify). [other]

A.4 Attitudes Toward Existing Solutions

How effective are the following developer tools in curbing toxicity? (5-point Likert: Do not exist, Ineffective - Very Effective)

1. *Text Restrictions* - Not allowing offensive text to be sent in chat

2. *Automated censoring* - Automatically removing or censoring offensive text in chat

3. *Muting* - Allowing players to select what text they see in chat functions

4. *Blocking* - Allowing players to block players that have previously proved toxic or unfriendly.

5. *Priming* - Developer tooltips that encourage prosocial or nontoxic behavior

6. *Honoring* - Allowing players to alert moderators to honorable or desired behavior by other players.

7. *Warnings* - Moderator- issued warnings for toxic behavior before a punishment is issued.

8. *Bans* - Moderator bans, preventing players from playing for a length of time

9. *EULAs* - Contract documenting the rights and restrictions applying to the game's players

10. *Codes of Conduct* - Ethical standards set by the developer

A.5 Issues with Existing Tools and Potential Solutions

[If ineffective or somewhat effective are chosen for any above]

If you feel any of the tools are ineffective, please explain why. How can they be improved? [Free Response]

B Descriptive Statistics

	mean	sd	median	min	max	range	skew	kurtosis	se
Age	19.87	1.77	20.00	14	25.00	11.0	0.43	0.50	0.07
Gender	1.80	0.59	2.00	1	4	3	0.57	1.97	0.02
Ethnicity	1.81	1.15	1.00	1	4	3	0.89	-0.92	0.05
B Embarrassment	3.49	0.86	4	1	5	4	-0.34	0.11	0.04
B Name Calling	3.78	0.85	4	1	5	4	-0.35	-0.21	0.04
B Griefing	3.86	0.81	4	1	5	4	-0.47	0.27	0.03
B Harassment	3.47	1.01	4	1	5	4	-0.28	-0.5	0.04
B Identity Based Harassment	2.9	1.25	3	1	5	4	0.09	-1.01	0.05
B Threatening	2.95	1.21	3	1	5	4	0.04	-1	0.05
B Doxxing	1.68	1.05	1	1	5	4	1.57	1.68	0.04
B Stalking	1.48	0.95	1	1	5	4	2.13	3.79	0.04
B Swating	1.35	0.84	1	1	5	4	2.61	6.39	0.04
B Experience	2.32	0.7	2.16	1	5	4	1.15	1.75	0.03
V Embarrassment	2.54	0.96	3	1	5	4	0.33	0.03	0.04
V Name Calling	2.74	1.04	3	1	5	4	0.14	-0.37	0.04
V Griefing	3.04	1.02	3	1	5	4	-0.08	-0.53	0.04
V Harassment	2.48	1.08	2	1	5	4	0.36	-0.44	0.05
V Identity Based Harassment	1.78	1.13	1	1	5	4	1.32	0.68	0.05
V Threatening	1.86	1.04	2	1	5	4	1.15	0.66	0.04
V Doxxing	1.16	0.56	1	1	5	4	4.23	19.51	0.02
V Stalking	1.15	0.57	1	1	5	4	4.47	21.34	0.02
V Swating	1.1	0.49	1	1	5	4	5.56	32.87	0.02
V Experience	1.65	0.53	1.51	1	5	4	1.95	6.43	0.02
Amount of Gameplay	15.41	13.26	10	0	100	100	1.96	5.78	0.56
Length of Gameplay	9.38	3.69	10	1	20	19	-0.03	-0.39	0.16
Gamer Identity	4	0.98	4	1	5	4	-0.9	0.41	0.04
Perceived Expertise	3.71	0.85	4	1	5	4	-0.56	0.21	0.04
Ranked Play	3.15	1.12	3	1	5	4	-0.12	-0.78	0.05
B Laugh	2.7	1.1	3	1	5	4	0.14	-0.64	0.05
B Retaliate	2.55	1.17	3	1	5	4	0.35	-0.68	0.05

	mean	sd	median	min	max	range	skew	kurtosis	se
B Ignore	3.2	1.06	3	1	5	4	-0.36	-0.43	0.05
B Mute	3.34	1.14	3	1	5	4	-0.41	-0.55	0.05
B Leave Chat	3.43	1.06	4	1	5	4	-0.52	-0.26	0.04
B Leave Match	1.99	1.07	2	1	5	4	0.84	-0.2	0.05
B Take a Break	2.2	1.07	2	1	5	4	0.54	-0.61	0.05
B Quit Indefinitely	1.57	0.86	1	1	5	4	1.46	1.54	0.04
B Refocus Chat	2.8	1.11	3	1	5	4	0.07	-0.69	0.05
B Support Victim	2.51	1.02	2	1	5	4	0.33	-0.34	0.04
B Call out Perpetrator	2.63	1.16	3	1	5	4	0.26	-0.73	0.05
B Report	3.21	1.36	3	1	5	4	-0.18	-1.18	0.06
V Laugh	2.94	1.15	3	1	5	4	-0.12	-0.81	0.05
V Retaliate	2.61	1.22	3	1	5	4	0.23	-0.92	0.05
V Ignore	3.36	1.16	4	1	5	4	-0.46	-0.51	0.05
V Mute	3.35	1.22	3	1	5	4	-0.4	-0.68	0.05
V Leave Chat	3.25	1.23	3	1	5	4	-0.32	-0.88	0.05
V Leave Match	1.93	1.11	2	1	5	4	0.99	0.05	0.05
V Take a Break	2.21	1.11	2	1	5	4	0.53	-0.66	0.05
V Quit Indefinitely	1.6	0.92	1	1	5	4	1.61	2.16	0.04
V Refocus Chat	2.6	1.19	3	1	5	4	0.18	-0.97	0.05
V Ask for Support	1.78	1.01	1	1	5	4	1.16	0.46	0.04
V Call out Perpetrator	2.51	1.27	2	1	5	4	0.36	-0.98	0.05
V Report	3.23	1.45	3	1	5	4	-0.21	-1.29	0.06
Text Restrict	2.09	0.91	2	1	4	3	0.50	-0.54	0.04
Censoring	1.09	0.89	2	1	4	3	0.70	-0.37	0.04
Muting	3.05	0.90	3	1	4	3	-0.67	-0.37	0.04
Blocking	3.30	0.81	3	1	4	3	-0.92	0.08	0.04
Reporting	2.53	1.02	3	1	4	3	-0.01	-1.13	0.04
Priming	1.67	0.86	1	1	4	3	1.08	0.23	0.04
Honoring	2.19	0.92	2	1	4	3	0.32	-0.77	0.04
Warning	2.33	0.89	2	1	4	3	0.08	-0.79	0.04
Ban	2.92	0.94	3	1	4	3	-0.47	-0.74	0.04
EULA	1.66	0.98	1	1	4	3	1.24	0.24	0.04
Code of Conduct	1.72	0.95	1	1	4	3	1.12	0.16	0.04