"I Don't Have Faith in the Developers to Use My Feedback": Understanding Player Values and Expectancy for Reporting Systems in Video Games

ANONYMOUS AUTHOR(S)

Reporting systems in multiplayer video games allow players to express their dissatisfaction with others and combat in-game toxicity. In this work, we examined the act of reporting through the lens of expectancy-value theory. Using a distributed survey (n = 98) and follow-up interviews (n = 19), we explored the value players place on reporting, their desired outcomes, and their expectations that these outcomes will be achieved. Our findings revealed that reporting is motivated by both altruistic and retributive factors, with players seeking short-term revenge while also looking to foster an improved long-term community. Yet, players felt that reporting may not always meet these goals, with belief in the system being mediated by factors such as developer reputation, reporting transparency, and alignment with the community. By understanding the value and expectancy of reporting systems, we discuss their implications on broader digital moderation and consider current and potential future designs of reporting systems.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: video games, reporting systems, flagging, multiplayer games, moderation

1 Introduction

Understanding and handling toxic encounters within games has long been a challenging area of ludological research. Toxic encounters can manifest in communicative forms, such as insults in voice or text chat, or non-communicative forms, such as intentionally ruining a game through purposefully throwing or quitting prematurely [27, 34, 39]. Toxicity is important for game developers to manage, as it can ruin players' experiences as well as propagate negativity, frustration, and vitriol [23, 33, 87]. Yet within certain gaming communities, toxic behaviour has unfortunately become increasingly normalized, with players viewing it as inherent to a competitive mindset [5, 78]. The primary feature offered to players to deal with toxic encounters is through in-game reporting (or flagging) systems [36, 60]. After a player submits a report regarding a perceived toxic player, it is sent to review — this system ends up deciding whether a punishment is warranted, and if so, how severe it should be. Although the goal of reporting systems is to combat toxicity, their effectiveness and practical usage remain lingering questions [26, 36, 41, 60].

For the player who initiated the report, the extent of how they can combat a toxic encounter within a game largely stops and ends at reporting. However, the effects of the toxic encounter can

extend long after. Research has shown that dealing with toxicity can affect a player's long- and short-term mood and self-esteem [23]. Yet, beyond these negative factors, when players submit a report, they hold onto expectations — a desire for accountability and consequences. We address the question of *what outcomes and expectations do players expect to get out of their reports*, especially in an *ephemeral* context like games, where it can be rare to re-encounter the same toxic individual.

We view expectations and outcomes as motivations for action through an **expectancy-value theory (EVT)** lens [84, 85], which has not yet been studied in the context of reporting systems in games. EVT frames actions as being motivated through a balance of value (the desire for specific outcomes) and expectancy (the likelihood or belief in meeting such outcomes); little has been done on whether reporting systems serve their intended purposes from the perception of *what players expect*. We ground this gap in our present study around the following research questions:

- **RQ1**: What values do players hold that are associated with reporting in video games?
- **RQ2**: To what extent does player perception of the consequences of reporting align with their desired outcomes?
- **RQ3**: What factors shape players' expectancy of their desired outcomes regarding reporting?
- **RQ4**: How do the prior factors play a role in broader perceptions of trust and expectations regarding the game?

RQ1 and RQ2 focus on values, as well as consider how players reconcile potential differences between their expectations and the outcomes that occur; RQ3 and RQ4 focus on expectancy, considering how likely users are to believe a reporting system is to provide value and how these factors influence broader perceptions regarding the game itself, the developers, and the community.

To address these questions, we employed a mixed-methods approach. To first explore the space of reporting motivations, we created and widely distributed a formative survey to gain broad insight into player experiences and expectations. We then conducted follow-up interviews to gain a deeper understanding of players' introspective feelings and motivations for reporting. The conceptual framework of EVT shaped our analysis of the data, as we discovered that participants shared values related to both retributive and altruistic reporting motivations. Yet, players often doubted that reporting systems can fundamentally meet their expected outcomes; notwithstanding that these outcomes primarily rely on a basis of trust, which is not always present. By examining the motivations behind reporting, the primary way of addressing toxicity, we broadly understand how people expect toxic behaviour should be identified and punished, and what the desirable outcomes are — for themselves, for the community, and for the reported perpetrator.

2 Related Works

We first examine ludological research on flagging and toxic encounters to understand reporting behaviours in gaming contexts. We then expand this perspective by considering reporting within broader governance, both in real life and in online contexts. Finally, we explore the theoretical framework of EVT and its relevance to interpreting player motivations and expectations.

2.1 Toxicity and Reporting in Game Contexts

Reporting in games is the primary way for players to combat toxicity, the latter being an enduring obstacle for both developers and researchers alike. The term toxicity can be broadly understood as the various negative behaviours that involve communicative abuse or disruptive gameplay behaviour [5, 33, 34], and has often been studied in team-based multiplayer competitive games such as *League of Legends* or *Overwatch* [27, 34, 36, 65]. Toxic behaviours can manifest in many ways; Kou et al. separated them into buckets of communicating aggression, cheating, hostage holding (keeping others in a negative situation), mediocriticizing (not trying hard to win), and sabotaging

(consciously lowering a team's chances to win) [34]. In multiplayer games, the competitive nature and the desire for personal ranking can create frustration towards teammates; the feeling of powerlessness or control over the gameplay outcome further induces toxicity [34, 95]. Factors such as game genre, player interactions, and communication channels through text or voice can directly instigate toxicity within individual matches [27, 55]. Personal factors such as gender identity can also affect the perception and manifestation of toxicity in games [92].

Although game developers often have community standards to highlight toxic actions by stating what actions are permissible or impermissible, *player perception* of toxicity is subjective and contextual. Laumann [42] and Laato et al. [39] showed that different actions are perceived differently in terms of severity, based on the underlying context and the frequency of the negative behaviour. Perception of toxicity is also affected by the broader attitudes within the gaming subculture. Beres et al. [5] and Türkay et al. [78] suggest that a level of negative behaviour has become normalized in gaming, with people viewing it as simply part of the competitive environment, to such an extent that players may even view positive communication as sarcastic rather than authentic [58]. Regardless of the accepted level of toxicity, being the victim of a perceived toxic encounter within games can feel frustrating.

From the opposing perspective, researchers have investigated the reasons why players might grief and exhibit toxicity within games. These motivations often revolve around personal pleasure, for instance, by exhibiting a form of power and control over others [1, 30]. From a self-determination theory perspective, Achterbosch et al. showed that griefing can increase the feelings of autonomy and competence of the perpetrators [2]. Researchers have considered ways to handle toxicity [87, 88]. For instance, Reid et al. developed in-game tools to support the victims of toxicity that increased their feelings of control and offered support [65]. However, players often leaned towards familiar in-game tools such as muting and reporting, as they offered actionable solutions [65].

Reporting systems are present in many online games (and digital platforms) [36, 60]. Ideally, users would report toxic behaviour that breaks the game's guidelines; these reports would lead to further punishments such as suspensions or bans [35]. However, while many players use reporting systems regularly [60], there remain issues regarding their perceptions and functional use. The standards for flagging differ from person to person, and unwarranted reporting of players who may not deserve it means that the system can be abused beyond its intended purpose [36, 60]. Players also question reporting's perceived effectiveness — despite using a system meant to deal with toxicity, players do not always observe that toxic players are punished or that negative behaviour has reduced [26, 36, 41, 60]. Certain systems have at times relied on crowdsourcing punishments based on reports through jury-based methods (e.g. *League of Legends*' Tribunal system); people were passionate about participation but could struggle with the system or act beyond their roles [37]. With more and more players, automated systems for toxicity detection and report handling have become popular in practice as well as in research [9, 12, 50, 91].

Contrasted against prior research, our present work investigates reporting (and its tie to toxicity) at a higher meta-level. Extending on the studied in-game experiences and immediate motivations for reporting, we link these experiences towards broader motivations, the impact of reporting outside of the singular encounter, and the expected outcomes of reporting compared and contrasted against player expectations. Kou et al. highlighted how reporting behaviour represents a user-participatory subset of a broader moderation process [36]. This wider lens allows us to understand reporting as a singular component of a more complex socio-technological process tied to community norms, player values, and system design.

2.2   Moderation, Punishment, and Approaches to Justice

People comply with the law for various reasons, including fear of punishment [79] and personal morality [57]. Research has shown that individuals are more likely to obey the law when they perceive it to align with their personal beliefs [28]. However, when a law is perceived as lacking legitimacy or moral justification, the mere threat of punishment is insufficient to ensure public compliance [16]. Low legitimacy can undermine people's trust and confidence in institutions such as the police or courts, leading to increased law-breaking behaviour and reluctance to cooperate with law enforcement [28]. Similarly, a lack of transparency in the legal process — such as unfair treatment by legal authorities — can also discourage cooperation [8]. In contrast, when people feel they are treated equitably under the law, they are more likely to comply with it [15]. One way of being complicit with the law is by reporting infractions — in addition to self-regulation, people also report illegal behaviour. Reporting is mediated by the misconduct's severity and encouraged by social integrity [32] — people are more likely to report misconduct when they perceive it as highly severe and in a positive ethical climate. Such environments encourage transparency and accountability, supporting individuals in taking ethical actions [32].

To draw comparisons with other types of governance, we also examine moderation practices on non-game digital platforms, such as social media and online forums. Content moderation on these platforms plays a crucial role in addressing issues like toxicity, which can significantly impact user engagement [4]. Content moderation refers to the process implemented by platform managers to oversee and filter user-generated content to protect users from harmful materials, such as misinformation, disinformation, hate speech, and online extremism [66]. This process involves determining which posts should be removed and which users should face restrictions [66]. In discussions surrounding content moderation, a critical tension arises between upholding freedom of expression and preventing harm caused by misinformation [38]. This moral dilemma has been explored in prior research, which found that most people prioritize harm prevention over protecting free speech [38]. This preference may stem from concerns about the negative impacts of misinformation on societal well-being.

While automated tools play an essential role in content moderation, user reporting remains a critical component [51]. Users report others for various reasons, including perceived violations of platform rules or personal grievances [21]. While user reporting serves as a vital mechanism for flagging potential issues, its effectiveness and fairness are closely tied to broader principles of transparency and accountability, such as those outlined in the Santa Clara Principles [70]. In 2018, several organizations, advocates, and academics launched and endorsed the Santa Clara Principles, which are a set of guidelines focused on transparency and accountability in online content moderation to support and empower users [67, 70]. The guideline included several core elements, including notices, explanations and appeal mechanisms — platforms must inform users when their content is removed or restricted, clearly state which rule was violated and how it applies to the removed content, and provide users with a fair and transparent way to appeal moderation decisions. We extend upon these principles while focusing specifically on user reporting as a mechanism for moderation and its implications for governance on multiplayer gaming platforms.

2.3   Theory of Motivations - Expectancy-Value Theory

There are many different frameworks of motivations that have been explored in academic ludology. For example, self-determination theory, which buckets psychological needs into those satisfying autonomy, competency, and relatedness, has often been used as the lens to investigate why people are attracted to gaming [18]. Ryan et al. show that such feelings of autonomy and competence during gameplay are associated with personal enjoyment and preference [69]. However, whereas

self-determination theory explains psychological motivations behind a sustained activity over time, we considered expectancy-value theory (EVT) as a suitable framework for investigating instantaneous, one-off decision-making processes such as reporting.

Under EVT, motivation to perform a task is mediated by the concepts of *expectancy* and *value* [84–86]. Value can be broadly construed as the desire for specific end states, and Wigfield and Eccles' work has categorized values into those of *attainment value* — the importance of doing well on the task, *intrinsic value* — the enjoyment one gets from the task, and *utility value* — the importance of the task for the future [84, 85]. Tasks also instill a cost that detracts from value, which comprises the time and effort of the task, the external effort to advance the task, opportunity cost for alternatives, and changes in emotional affect [19, 22]. Expectancy refers to the expectations, or *belief*, that behaviour can lead to the desired outcomes [19, 85]. EVT has largely been employed in educational and developmental contexts to understand the motivations for achievement [54, 62, 86]. However, it has been studied more broadly as well, such as in game contexts. Rachmatullah et al. found that outcome-expectancy belief was a positive predictor of flow experiences in games, whereas lower self-efficacy (personal belief) resulted in increased feelings of frustration [63]; Lin et al. highlighted the importance of achievement value on player engagement in massively multiplayer online role-playing game (MMORPGs) [44].

EVT provided a suitable framework for our work because our research questions strongly focus on expectations and outcomes — we explore factors such as what values people place on the outcomes when deciding to report in games, mediated by the likelihood of such consequences. Our work explores how the mental calculations core to EVT are affected by discrepancies between expectations and observed experiences, transparency of the process, and the alignment of norms.

3 Methods

To investigate our research questions, we performed a mixed-methods approach to data collection and analysis. We generally followed an explanatory sequential design [17] — starting with an early-stage, formative exploration of quantitative trends that informed deeper qualitative analysis to understand these trends. We started broadly, deploying a survey to garner broad initial sentiments regarding motivations and expectations behind reporting practices. Motivated by the survey results, we dug deeper into specific experiences and feelings through rich interviews with video game players. Our methods were reviewed and approved by our institute's ethical review board.

3.1 Survey

3.1.1 Survey Development and Procedure. The goal of the survey was to broadly explore the range of reporting habits in games, bringing together a wide range of participants from a wide range of games. We used this data to understand patterns in motivations for reporting, as well as people's general expectations for its outcome. The initial design of the survey was motivated by prior research and researchers' brainstorming — taking the view of researchers as experts, as all researchers had extensive experience with playing video games and using their reporting systems. Although our survey touched on questions later interpretable through EVT, such as the value of reporting (e.g. community-improvement, avoiding toxic players) and its potential expectancy of reporting (e.g. chances of consequence), the theoretical lens of EVT was not imposed during survey construction. Rather, we wanted to broadly understand the space and the most general trends around reporting behaviours.

To highlight these trends and to provide a level of explainability around the behaviours, the survey consisted of a mix of multiple-choice, Likert-scale and open-ended questions. The required Likert-scale questions asked participants to rate their agreement with statements regarding reporting

expectations, motivations, and outcomes; the optional open-ended questions mainly offered space for participants to freely expand upon their ratings with their rationale.

To begin, we asked participants to select behaviours they would report for and what punishment they expect for each behaviour, with the toxic behaviours inspired by Kou's prior classification [34], and the punishment outcomes based on our own gameplay experiences combined with Ma's classification of punishment types [47]. To understand experiences around reporting, we asked participants to rate their agreement with motivations for reporting around various mediating factors, e.g. understanding of transparency, mental model of consequences, and potential for reformation. These were inspired by the researcher's readings of prior work in both game moderation [35, 36, 47] and broader moderation in social media [48, 94]. For instance, Zhang et al. [94] highlighted how participants may have differing understandings and motivations for reporting on social media, and we translated these questions into a games context. Zhang et al. [94] also raised how reporting is both a collective duty and an individual right, which became questions regarding the extent to which reporting in games is driven by collectivist outcomes or personal desires. Ma et al. [47] highlighted how transparency and fairness impact moderation in gaming; we asked questions regarding the extent to which punishment transparency might also play a role for the reporter.

The full survey is in the supplementary materials. Internal face validation was conducted through pilot testing — we asked HCI researchers with game expertise to subjectively critique our survey structure, identify confusing questions, and assess coverage of our RQs. We estimated that the survey would take 30 minutes to complete, and we offered compensation of $6 CAD.

3.1.2 Survey Participants. We distributed the survey online through Prolific[1], an online crowd-sourcing platform for user studies. In the survey posting, we highlighted that the main inclusion criteria were to have experience with video game reporting systems (which we also added as a screening question in the survey itself). We also pre-screened participants based on their approval rating and the number of previous submissions, aiming for higher-quality data. However, after data collection, we further cleaned the responses to remove responses with implausible fast completion times (under 4 minutes), incomplete surveys, and non-English responses, arriving at a final sample of 98 participants. The ages ranged from 18 to 62 (average: 28.5), and the gender distribution was 62 male, 35 female, and 1 non-binary. In terms of gaming experience, 62 participants had played games for over 10 years, 25 had played games for 5–10 years, 6 had played games for 2–5 years, and 5 had only played games for less than 2 years. In terms of frequency, 30 participants played games for more than 10 hours a week, 34 participants played 5–10 hours, 21 played 2–5 hours, 11 played less than 2 hours, and 2 participants indicated that they currently are no longer playing games. We found the demographics of our survey participants to be acceptable, providing a satisfactory sample of gaming experiences and frequencies from different age ranges.

3.2 Interview

3.2.1 Interview Protocol Procedure. The interviews chronologically followed the surveys, after we had already started initial data analysis and highlighted initial patterns. The goal of the interview was to delve deeply into the experiences and rationale underlying the trends in motivation and expectations from the survey. The interview provided a deeper explanation and reasoning for these trends, as well as exceptions and counterexamples. The interviews asked about people's reporting experiences, what about the encounter stuck with them, and the relationship between reporting and community improvement.

The qualitative, flexible nature of the semi-structured interview allowed us to understand experiences at a broader scope compared to the questionnaire — for instance, instead of ranking agreement

---

[1]https://www.prolific.com/

with existing reasons for reporting, we could ask participants to provide their existing rationale unbounded by our prior expectations. We view the interview as still exploratory — exploring the space of reporting habits, their experiences, and their expected outcomes. In this spirit, our interview questions were related to EVT constructs but still open-ended, aiming to understand their rationale for reporting and what kind of outcomes they expected and why (e.g. "Why do you report people?", "What is reporting meant to do?", etc.), even though follow-up questions could diverge into broader exploration. We conducted initial pilot testing with HCI researchers to evaluate the subjective face validity of our interview questions.

The interview protocol can be found in the supplementary material. The interviews took approximately 40–60 minutes online over Zoom, and each participant was reimbursed $16 CAD. Audio recordings were collected with participant consent.

3.2.2 Interview Participants. Participants were recruited from a listing on our institute's paid studies page, with the main inclusion criteria being having experience with reporting others in games (in addition to being fluent in English). We recruited a sample of 19 participants spanning a wide variety of different genres, gaming experiences, and reporting habits — some participants had only recently started playing games and had only reported toxicity a couple of times, others had been playing multiplayer games for over a decade and issued a multitude of reports. The researchers subjectively felt that this sample of participants provided sufficient information power [49], based on factors including the specificity of experience, the quality of interviews, and so forth; this sample size also exceeds local standards for HCI [11]. Information power was subjectively judged by the primary researcher on an ongoing basis during the interviews (in discussions with the research team); we stopped collection when we found that additional data were yielding diminishing new insights into answering the research questions.

One factor that might cause people's experiences to differ is the genre of games that people play and their experience with games as a whole. As such, before the interview, participants were asked to send the researcher information about the games they had played and reported people in (the researcher read up on the game in the cases where they were not familiar), their frequencies of reporting, and whether they had been punished based on other reports. We kept the latter two questions unstructured so participants could self-report their diverse and unique experiences, and participant responses to them can be found in the work's supplemental material.

Furthermore, during the study, we collected information regarding participants' age, gender, frequency of video game play, and their years of experience playing games (which we summarized in Table 1). The average age of the participants was 24.3, the gender distribution was 9 male, 9 female, 1 non-binary, and most of the participants were experienced gamers with over 10 years of experience. Although a wide range of games were mentioned, the majority of games discussed in-depth were team-based multiplayer games (such as *League of Legends*, *Overwatch*, *Valorant*).

3.3 Data Analysis

The data analysis process focused on the concept of triangulation [13] — using multiple methods to understand the reporting phenomena. We first aimed to understand reporting phenomena broadly, before narrowing into deeper interpretation. We started with our collected survey data, which consisted of both quantitative and qualitative data. Our quantitative data was largely exploratory — we designed the questions around understanding trends in the motivations and expectations around reporting patterns. We inspected the quantitative data through exploratory data analysis to understand these trends at a high level. However, given the survey's formative nature, we did not conduct any specific statistical tests on the quantitative data.

Table 1. Summary of Interview Participants

ID	Age	Gender	Games Discussed	Frequency of Play (Weekly)	Years of Experience
P1	21	M	World of Tanks Blitz, World of Warships, Clash of Clans, Battlefield 2	<2 hours	>10 years
P2	27	M	League of Legends	5-10 hours	>10 years
P3	29	M	League of Legends, Overwatch	2-5 hours	2-5 years
P4	25	M	League of Legends, Overwatch, Fortnite	2-5 hours	5-10 years
P5	30	F	Teamfight Tactics, Clash Royale	>10 hours	>10 years
P6	23	F	League of Legends, Valorant, Club Penguin	5-10 hours	2-5 years
P7	22	M	Valorant, Rainbow 6 Siege, GTA Online, Chess.com	2-5 hours	5-10 years
P8	20	F	Golf Battle	5-10 hours	>10 years
P9	21	NB	Minecraft, Roblox, Identity V	2-5 hours	>10 years
P10	19	M	Valorant, Overwatch, Fortnite	5-10 hours	>10 years
P11	18	M	Rainbow 6 Siege, PUBG, Rocket League, Counter-Strike 2	>10 hours	2-5 years
P12	23	M	Call of Duty Mobile	5-10 hours	>10 years
P13	20	F	Fortnite, Splatoon	5-10 hours	>10 years
P14	38	M	Dota 2	>10 hours	>10 years
P15	18	F	Valorant, Roblox, PUBG	<2 hours	>10 years
P16	27	F	Deadlock	5-10 hours	<2 years
P17	32	F	Overwatch	5-10 hours	>10 years
P18	23	F	Valorant, League of Legends, Counter-Strike 2	>10 hours	5-10 years
P19	25	F	Valorant	5-10 hours	5-10 years

Still, we used patterns from our quantitative exploration to inform our qualitative analysis of the open-ended survey responses, employing a thematic analysis approach [10]. The first and second authors collaboratively began with initial coding on the data, generating code labels that were iteratively refined as the process continued. Visual mapping approaches were used to develop relationships and categorize the codes, which later informed the development of themes. Although the coding was done independently, both authors helped generate and discuss code labels, relationships, and themes. These themes formed initial inductive trends regarding reporting

practices. Even though our collected data from the survey was broadly related to constructs of motivations, value, and expectancy, it was during data analysis that the theoretical lens of EVT sharpened. When analyzing the data, we found that EVT provided a rational framework for us to view and articulate why people were motivated to report (i.e. the value from reporting), and what kind of consequences people wanted from reporting (i.e. the expectancy of reporting).

Consequently, the EVT framework helped shape both the structure and analysis of our interviews. Although the questions in our interview were influenced by our findings from our survey, the analysis of the interviews began with an initially inductive approach, understanding the data as a whole without prior assumptions. This was done to check whether there were findings that emerged beyond the initial questionnaire findings or the broader theoretical EVT framework. The first and second authors again started with initial exploratory coding to understand the experiences and perspectives of the participants. In this initial exploratory coding, and ongoing throughout the analysis process, we engaged in ongoing discussion about the data, the data labels, and their own thoughts and interpretations of the data (which was also discussed regularly with the team).

As the analysis progressed, we refined and organized the initial codes. This process of refinement was iterative, driven through comparisons across participant data and discussion among the researchers when codes were too granular or too broad to be effectively applied. After this coding process, we shifted towards a more deductive, theory-driven perspective to translate the relationships in our code labels into defined themes. EVT informed the mapping and grouping of the codes, as we developed relationships and hierarchies in the data (see supplementary material); this process later crystallized into themes around the "expectancy" and "value" of reporting. For instance, the codes "Reporting for Spite / Revenge" and "Reporting for Altruistic Reasons" highlight the split motivations that interview participants spotlighted, which became grouped under the EVT construct of "People's Values". On the other hand, codes like "Feeling like the Developer Cares" and "Communication of Reporting Feedback" tied to how participants shaped consequences for their reports, and became grouped under the EVT construct of "People's Expectancy".

As our interview data extends our survey data through richness and depth, we constantly made links between the two sources as we progressed deeper into the analysis. We considered how patterns from the interview data could both fit into and challenge existing trends we had highlighted previously. Given the rather formative and early-stage approach of our questionnaire (especially as not everyone opted to answer every open-ended question), we mainly lean on qualitative quotes from our rich interviews in the findings. The survey data follows similar trends, and we use the quantitative survey data as support in presenting our findings.

4 Findings

4.1 RQ1: What values do players hold that are associated with reporting?

We outline the values regarding reporting — subjective task value is constructed by potential benefits weighed against costs [19, 53]. People report for different gameplay reasons (illustrated in Figure 1, with grouping inspired by Kou [34]). However, how people's perceived task *values* for reporting are different depending on their motivation, which we initially explored through the questionnaire responses (see Table 2) but expanded on during our interviews. We identified two main categories of motivations driving distinct benefits: **retributive** motivations — of wanting to 'get back' at the reported player and **altruistic** motivations — of wanting to improve the community for everyone. Some of the interview participants were more readily influenced by one set of such motivators in reporting, but the vast majority indicated a mixture of both.

*4.1.1 Benefits Associated with **Retributive** Motivations.* Participants discussed how part of the motivation that comes with reporting is retributive — wanting to take their anger out and blame
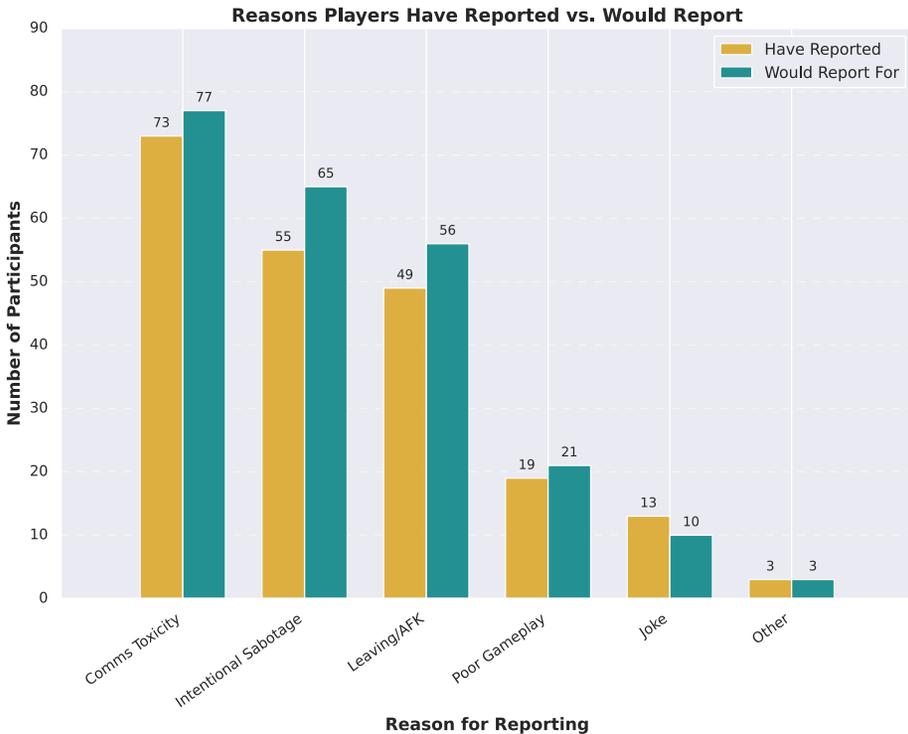
Fig. 1. Number of questionnaire respondents that have reported and would report for specific reasons (out of 98 respondents total)

someone after having a negative experience. We interpreted this motivation to be highly influenced by the player's own temporary experience within the game, such as the match outcome or the player's temporary mood. For instance, P2 mentioned that sometimes after losing a match in *League of Legends*, players might feel the need to:

> *"Let out frustration ... you're looking for someone to blame almost, I would say, especially right after you've lost"* - P2

Retributive motivations tied rather closely with immediate intrinsic value — players used reporting as a way to feel better in the short-term, as doing so could provide *"a sense of catharsis in hitting the button"* (P14) and *"revenge ... you've been toxic to me so I'm going to report you so you can't speak next game"* (P19). The reporting system provides the primary way for a player to vent out their negative emotions, as well as a way of regaining control. In multiplayer team games, players felt a lack of control as they were forced to engage with this negative player for the full duration of the game — *"I feel like toxicity in that context you can't really control ... because League of Legends each game is 30-35 minutes"* (P2). Through reporting, players regained this control — P7 states that the reporting player exerts *"some amount of power"* because of *"the ability to get someone penalized"*.

Penalties tie into the utility value behind retributive motivations — the outcome that people desire from reporting from this perspective is the punishment of the reported player. In the ephemeral context of games, where it can be rare to run across the same players in most ranks of gameplay, we interpreted this actual outcome as rarely offering any actual long-term self-benefit — whether

Table 2. Agreement levels to the questions regarding motivations for reporting from 1 (not a motivation at all) to 5 (a strong motivation) from our questionnaire [n=98]

Question ID	Reason	Mean	Std. Dev
Q1	To make the game a better place for everyone	4.59	0.82
Q2	To prevent future toxic behaviour and improve the quality of gameplay for others	4.62	0.68
Q3	To feel a sense of self-satisfaction or closure after experiencing negative behaviour	3.10	1.40
Q4	To avoid encountering negative behaviour in future games for myself	4.43	0.85
Q5	Because it is part of my responsibility as a player to uphold the community standards	3.69	1.20
Q6	To hold players accountable for their actions	4.36	0.94

someone you never see again is punished or not does not truly matter. P2 extended on their prior statement, revealing that *"deep down, you know that it's not really a rational thing that you're doing"*.

Since retributive motivations are impacted by mood and outcome, reports made with such motivations may not always have the best intentions. Most participants spoke out against reporting people for invalid reasons, such as *"if someone has a better KDA [author note: KDA = Kills/Deaths/Assists, a common metric for performance in multiplayer games] than them"* (P15). At the same time, several participants mentioned that they sometimes do use the report button for reasons that admittedly may not be deserved:

> *"Sometimes I used to report people for fun, if I lost or something"* - P12

> *"If I get tilted and then I see someone's not doing too well even though they might just be not having a good day"* - P10

Thus, even though reporting is primarily viewed as a way to combat toxicity, it can be used in a toxic manner as well — potentially ruining someone else's experience due to personal frustrations.

*4.1.2 Benefits Associated with **Altruistic** Motivations.* The other main motivator associated with reporting was altruism — wanting to improve the gameplay experience and the broader community for everyone. This motivation was influenced by the player's personal values, their perceptions of accountability, and their moral judgment. P9 stated the role of reporting was for *"making it better for the next player"* and P6 ties it towards *"fair[ness] and justice"*. Reporting behaviour in this case was mediated by the perceived intention of the other player towards not just themselves, but towards others as well. From this perspective, the reporting system serves as the primary way to combat negative behaviour and make a difference in the community. P18 highlighted their experience with sexist comments in video games, reporting to help others in the future:

> *"There's not a lot of girls that play ... so when I report someone, if a girl starts playing ... she doesn't have to experience that"* - P18

We associate this idea of wanting a better community with the utility value of altruistic reporting. Participants were aware that one report does not change the community. Thus, in some cases, there is no immediate altruistic intrinsic value for the players themselves — as P13 mentioned *"talk and*

things like that, they don't personally bother me that much, I'll normally just mute and move on", highlighting how, for their immediate experience, dealing with the minor annoyance is simple even without reporting. However, P13 continued — *"I feel like it's my responsibility to report actions ... that there's not some younger girl on the other side of the screen that will experience that and have that affect them"*.

For most participants, the core altruistic value for reporting is in cultivating a positive environment for the future, beyond the ephemeral, toxic encounter. This is not solely for the benefit of others, as the reporting player also wants to experience an improved, less-toxic environment. This tied into the personal value that certain players saw in games as a whole:

> *"Gaming is to build a community of people you would probably game with all the time ... being rude towards somebody ... takes away from the meaning of it being a fun thing"* - P16

4.1.3 Benefits are Mediated by the Dimensions of Specific Games. The benefits of reporting are unique to specific game contexts; several factors affected how players perceived reporting value within specific games. Two factors discussed previously were the ephemerality of interaction — where it is rare to see the same player in different matches in multiplayer team-based games, and the forced temporal interaction — where players are forced to share the gaming space with others for a set amount of time. However, participants also brought up how factors such as team dynamics and cooperation requirements that could induce toxic reportable behaviour, e.g. *"when someone isn't playing with the team then that makes me want to report them"* (P10); this was exacerbated by competitive settings *"in competitive games, the frustration is usually higher ... so reporting and other stuff is more [often]"* (P11). Communication channels within games were also discussed, especially regarding the impact of voice chat: *"it makes people more toxic because it's easier to talk shit to your teammates"* (P10). When considering the list of discussed games in Table 1, there is some potential bias towards multiplayer team-based games. However, people's responses suggest that reporting is more prevalent in these types of games precisely because of these contextual reasons.

4.1.4 Costs Associated with Reporting. Under EVT, task value increases through the benefits of an action but decreases as a result of its costs [19]. In the context of reporting, however, the *cost* associated with reporting was generally perceived to be low. Participants indicated how games typically make it extremely easy and low effort to put in a baseline report, making it accessible to any person — *"you just have to click a couple buttons"* (P10). This relatively low cost of reporting also aligns with our survey response for the question *"it requires a great deal of effort for me to report a player"* (mean=2.84, std=1.39).

A reporting system that would incur more effort and a higher cost might be less attractive for people to use, e.g. *"if I had to manually go to a website or form I would probably not report almost anyone"* (P17). However, a low cost can also be a problem given our prior discussion regarding reporting as a potentially abusable behaviour — P2 mentioned that it is precisely because of this low cost that they see it as *"kind of a joke just because you can just do it with a few buddies with no real evidence needed"*.

Even though the baseline cost for submitting a report is low, games often provide the opportunity to add text as supporting evidence. Some participants were willing to incur this added effort to obtain the aforementioned value, e.g. P18 mentioned that with a *"free text option I'd be writing to the max characters"*. Tying into the retributive rationale for reporting, P2 mentioned that the ease of reporting means that they can easily attain intrinsic satisfaction during the game, *"just take a pause to report the person and it's like a minor cathartic moment"* (P19).

4.2 RQ2: To what extent does player perception of actual consequences of reporting align with expected outcomes?

Building upon participants' expectations of reporting outcomes, which differed based on their underlying motivation, we examine the extent to which players believed reporting could lead to these desirable outcomes. Specifically, we consider whether players perceive reporting as an effective means to (1) punish offenders and (2) foster a better community.

4.2.1 Reporting and the Expectation for Player Punishment. The retributive outcome associated with reporting is for the reported player to be punished. However, all players were cognizant that not every one of their reports led to the desired effective punishment against the player. Other than the most egregious or easily detected cases, participants indicated how game moderation systems typically punish reported players after a sustained pattern of wrongdoing. Participants had mixed attitudes towards this, with some being accepting of this leniency — *"maybe they should be given a chance"* (P5). However, others felt more strongly about punishment after a single offence:

> *"If you grief one [game], you're still harming a lot of people's experiences"* - P10

> *"Letting people get away with it once because they've never done it ... kinda defeats the purpose of the system"* - P2

For these cases, people were focused on their report as correctly identifying toxicity, so in that case, reporting *should* logically lead to punishment. Yet, despite their internal beliefs, we have already established that this cannot hold for every single report. We have acknowledged that the report button can be used for the wrong reasons, highlighting the importance of added contextual information — *"one report against a player who was never reported before could be for something severe, or it could not be"* (P16). Although participants *speculated* that not all reports were punished, the lack of transparency in typical report systems makes it hard for them to *confirm* — we extend upon reasons on why participants held such beliefs when we discuss outcome *expectancy* in the next findings section.

Even when the reported players were punished, players were ambivalent about the structural factors of the punishment. Several participants indicated that they would like to see more severe punishments to act as a deterrent; the survey responses also provided loose support to this: *"I believe that current punishments for my reports are too light"* (mean=3.26, std=1.06). This was mainly tied to the ease of circumventing the report:

> *"I don't think any of it is getting taken seriously ... people can just open a new e-mail address and a new gaming profile"* - P12

Returning to retributive values, players sought to punish the actual reported player (not their account) and attain revenge, but the ease of circumvention, especially in free-to-play games, makes the current consequence misaligned with the desired outcome. Players did note that harsher punishments could be employed to ensure the actual punished player is not allowed to continue playing, such as banning by IP address, banning based on hardware IDs, or account subscription fees, but those come with problems as well, e.g. *"people say IP banning could ban their siblings' account by accident"* (P5).

4.2.2 Reporting and the Expectation for a Better Community. The altruistic outcome associated with reporting is in fostering a better, less toxic community. As the short-term outcome of reporting is potential punishment, we begin by understanding what players consider to be the relationship between punishment and an improved community. Firstly, whether punishment was an effective way of quarantining toxicity was considered, however, we have already identified that the ease of making new accounts limits the effectiveness of punishment in this consequence, e.g. *"people get*

Table 3. Agreement levels to the questions regarding expectancy variables from 1 (Strongly disagree) to 5 (Strongly agree) [n=98]

Question ID	Statement	Mean	Std. Dev
Q7	My reports typically lead to the consequences I expect	3.32	0.88
Q8	I am typically satisfied with the outcomes of my reports	3.39	1.02
Q9	I am motivated to keep reporting players even if the outcomes do not always match my expectations	3.93	0.96

banned and make new accounts and ... continue the cycle all over again" (P18); in such a way, toxicity does not improve unless people are encouraged through punishment to change their behaviour.

Some participants were optimistic about punishment to deter toxicity and underscore change:

> *"Some people will be afraid of being temporarily banned or being punished. So at a point I think they will, they will behave themselves"* - P4

People may want to keep playing the game or have a value associated with their account that they do not want to lose, e.g. P11 shares an anecdote where *"one of my friends he's very toxic sometimes"*, yet after getting banned, *"he was forced to turn off team chat, so it kind of works sometimes"*. Thus, punishment can improve the community when people are incentivized to reflect on their actions and change, even if it is solely so that they can keep playing.

However, others were less optimistic about punishment as changing player attitudes — P2 states that games are *"not the kind of setting that people want to be reformed in"*, P10 extends that people's toxic habits only change if:

> *"The person already wants to change, but otherwise no, I don't think it'll help make that person a better person"* - P10

If people cannot reflect on their actions and change after punishment, then reporting does not help in making the community a better place, as the person simply returns and continues to act similarly. As P17 mentions *"I don't think [punishment] necessarily fixes or makes [the reported player] reflect"*. Despite this, players had difficulty identifying what alternatives could reach this outcome:

> *"Punishment isn't the best but you can't administer therapy for everyone ... punishment isn't a great deterrent and we know it doesn't really work. I wouldn't know what an alternative is"* - P18

This is partially due to factors beyond reporting (and thus out of the player's control), e.g. P6 mentions how, even as toxic players get punished, there is always an unending battle because *"there's always an influx of players"*.

4.3 RQ3: What factors shape players' expectancy of their desired outcomes?

We highlighted how the bridge between reporting action and resultant outcomes can be disconnected from the results from both retributive and altruistic perspectives. Here, we explore the factors that influenced why (or why not) players may **expect** their desired outcomes to be achieved. From Table 3, we find that survey respondents reported neutral-to-positive scores on direct expectancy (Q7) and outcome satisfaction (Q8). From Figure 2, factors such as transparency in process and results, clarity of reporting guidelines, consistency of perceived enforcement, and observations of

improvements in the game over time all comprise significant factors that comprise feelings of a player's *trust* in the system.

We found the concept of *trust* to be a major factor in affecting expectancy, appearing commonly in participant narratives and our data analysis. While trust is not the sole factor that impacts expectancy, it plays a major part in a reporting context simply due to the non-transparency of the system — players must hold a level of trust that the black-box reporting system operates correctly; as we find, this is mediated by trust in more tangible aspects, i.e. the developer and their fellow players. Trust is especially important in uncertain, opaque conditions, and can be shaped by personal views and experiences of external facets "around" an opaque system, such as the institution or company using that system [76]. We find that trust, along with other a priori construals, underpin the more EVT-aligned construct of individual *belief* [19] that reporting will provide meaningful value.
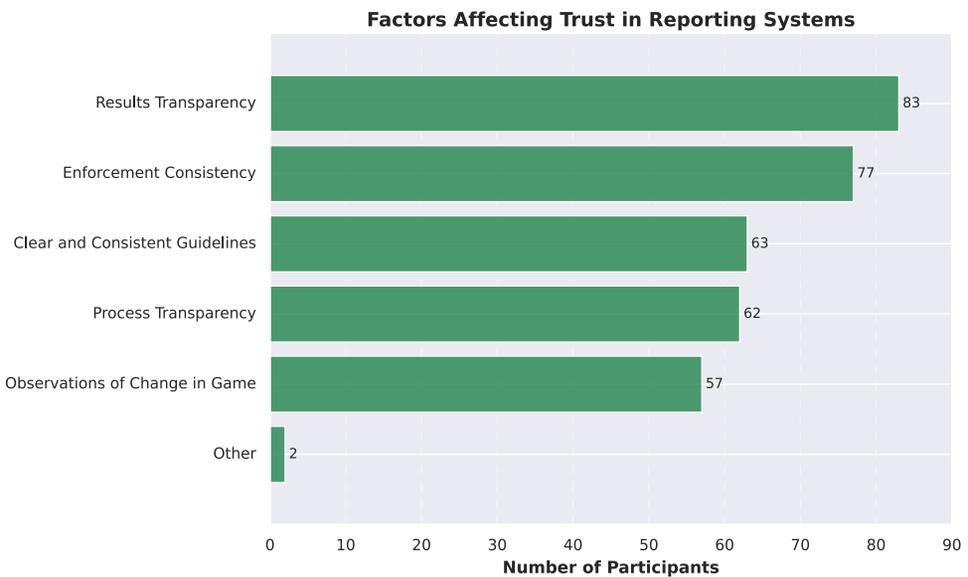
Factors Affecting Trust in Reporting Systems

Factor	Number of Participants
Results Transparency	83
Enforcement Consistency	77
Clear and Consistent Guidelines	63
Process Transparency	62
Observations of Change in Game	57
Other	2

Fig. 2. Number of questionnaire respondents that agreed with how much each factor affected trust (out of 98 respondents total)

4.3.1 Belief is Affected by Experiences. Participants' beliefs that their report would lead to a desired outcome were driven by their past and ongoing experiences. This comprised both previous knowledge of the developer and their approaches to toxicity, as well as the current perception of the effectiveness of specific reporting systems.

Prior Experience with Developers. The extent to which players perceived the developers' attitudes toward the game significantly influenced their expectancy related to reporting outcomes. If players trusted the game developers, they would hold a stronger expectancy of attaining desired outcomes. In contrast, when players believed addressing toxic in-game behaviour was a low priority for the developers, they had low expectancy of attaining their desired outcomes. Part of the trust in the developers came from the players' prior perception of reputation. P14 highlighted this perspective:

> "[Valve] [author note: Valve is a video game developer and publisher, notable for the Dota and Counter Strike series of games] is seen as a company that does things the right way

and have earned the trust of the community... it extends into all other facets of how they manage their games" - P14

One way of establishing this perceived trust and rapport is through visible community engagement:

"If a game developer is really involved with the community... it usually brings a lot of generally good people, positive people to the game as well" - P7

Developer actions play a major role in influencing their reputation, the perception of the community, and player engagement with the entire reporting ecosystem.

On the contrary, a low level of trust in developers can cause players to consequently fail to believe in the reporting system as having any value. For example, P8 had a negative perception of the developers of *Golf Battle*, stating that they prioritized short-term profits over creating a sustainable gaming environment:

"It's the case with many mobile games. The purpose is more about in-app purchases and profits for the company rather than making a successful long-term game" - P8

When players negatively perceive the game developers for other reasons, they subsequently distrust the game developers to handle reports as well — P8 expanded that *"You don't expect anything from the developer, and the developer doesn't really expect anything from the community either"* (P8). P2 shared a similar opinion, expressing that when players feel ignored by developers, they lose confidence in the developers' ability to address all issues effectively, stating that toxicity was:

"An issue that they could have really fixed if they wanted to a long time ago, but went neglected for so long that at this point it's too late" - P2

P2 cited how this loss of trust caused him to also lose belief in the efficacy of the reporting process *"you can't trust the developers that they're not punishing the right people or giving the right punishment for certain people"*. The loss of trust undermines player expectancy in the reporting process and contributes to a less supportive gaming environment.

Ongoing Experience with the Game. While dealing with toxicity, players were still engaged in an ongoing experience of playing the game, and one factor that shaped their expectancy was their observation of toxicity within the game over time. Whether the desired patterns of changes (of player punishment or community improvement) were actually observed fundamentally shaped expectancy, even if it was reliant on anecdotal evidence or was a result of the broader, more external ecosystem rather than solely reporting. For P13, seeing people complaining on Reddit or Discord about being banned actually *"ups my trust in the moderation system a bit more"*; for P14, expectancy was shaped through observed improvements in the broader gaming community:

"[Counter Strike] was an absolute cesspool back in the early-mid 200s... things have gotten so much better" - P14

On the other hand, P18 felt that *"nothing's really changed"*, which led them to not believe in the reporting system *"it genuinely doesn't matter... I feel like it just doesn't work"*. Such experiential factors shape beliefs on how attainable positive outcomes are, which are core to expectancy.

4.3.2 Belief is Affected by Normative Alignment. Participants' belief that reporting had value was driven by how well their own mental model aligned with both the reporting system and other players. Misalignment in the (subjective) definitions of punishable toxicity and the consequences of submitting a report could create inconsistent perceptions of enforcement, causing players to fail to expect the system to work as they believed.

Alignment on Reporting Purpose. Even though there are shared outcomes that players expect, player attitudes and perceptions towards how their actions lead to the outcome were varied. Some participants had a stronger view of their report as a direct signal of wrongdoing, where it logically follows that wrongdoing should be punished, and if nothing happens, then the system is broken. In such cases, reporting represents a near-causal mechanism — you report, and the system acts:

> *"If that person is not punished ...maybe there's something wrong with the system"* - P3

On the other hand, others had a more relaxed attitude towards reporting in that it acts as a single vote for toxicity within the broader scheme:

> *"It's creating a paper trail, it's creating an inaudible event that even if you know this person gets reported and nothing happens this one time, if other people run into toxic behaviours in the future from this person, it's more likely to get picked up on* - P14

These participants accepted that usual implementations of reporting systems mean that not everyone who behaves badly in their eyes will get punished, even though this acceptance can be begrudging at times — *"it's just like that's just how the system is"* (P9) and *"I kind of learned to not expect anything from it"* (P13). In these cases, reporting is a more probabilistic indication of wrongdoing. Overall, individual attitudes toward reporting purpose shape people's expectations as to whether reporting can meet their desired outcomes.

Alignment with Other Players. All participants believed the likelihood of punishment to be dependent on the number of reports levied on a perceived toxic player, where actions were more likely to be taken when multiple players reported someone, rather than just a single individual. As such, some players perceived the process of addressing and resolving the reported behaviour as a community effort rather than a purely individual one:

> *"It's not just about your individual report. It's more like a community-driven process, where members collectively vote on whether a person should be punished for toxic behaviour"* - P8

Thus, it is through the collective effort of multiple reports that the system can determine whether a player is consistently toxic or only exhibits toxic behaviour occasionally. This was perceived to help ensure fairness, particularly in the anonymous environment of online games:

> *"For some kind of fairness in the system, I think it makes sense for games to rely on a huge pool of complaints. So like a much larger pool than in real life"* - P1

Yet, not only does this misalign with the perception of reporting as a near-causal process, but it also means that an individual's belief in the reporting system to attain value is mediated by trust in the community to use the reporting system properly — as P17 stated *"If nobody reports, then nothing is gonna happen. Nothing is gonna change... but we can't force people"*. Yet several participants had low faith in other players to use the system properly. Even in a world where players who report are well-meaning, the community representation towards collective judgment on toxic behaviour is influenced by individual norms:

> *"Individuals perceive things differently in every aspect of life, including games. What feels toxic to me might not feel toxic to another"* - P3

Thus, while collective judgment has the benefit of helping mitigate abuse of the reporting system, the lack of a clear shared vision for players to judge toxicity and rule-breaking means that collective judgment can misalign with any individual's own reporting model and definitions of toxicity — *"it's kind of a loose system, it's a very emotion-based system"* (P2), which can cause the perceived outcomes to be inconsistent with individual expectations.

Table 4. Agreement levels to the questions on reporting feedback from 1 (Strongly disagree) to 5 (Strongly agree) [n=98]

Question ID	Statement	Mean	Std. Dev
Q10	It is important for me to know whether the player I have reported has faced punishment	3.69	1.05
Q11	It is important for me to know the player I have reported has evaded punishment	3.48	1.11
Q12	If they have evaded punishment despite my report, I feel like I would want to know why	3.79	1.15

4.3.3 Belief is Affected by Feedback and Transparency of Outcomes. Finally, participants' belief that reporting could lead to desired outcomes was affected by the transparency of the entire reporting-to-punishment process, and relatedly, whether feedback was provided after their reporting action. With more opaque processes, participants exercised a degree of guardedness that could lead them to question the efficacy of the system and whether their report even matters, eroding expectancy.

Presence of Reporting Feedback. When reporting systems consistently offered meaningful and timely feedback, players were more likely to believe in the process and continue reporting toxic behaviour. Such feedback made them feel like their efforts contributed to a healthier community, reinforcing that desired outcomes were being met:

> *"Sometimes you do get feedback that a player you reported recently was punished. I guess that's like, a lot of people do find that satisfying"* - P5

The quantitative survey results in Table 4 also provided support for this, as respondents indicated desires to receive feedback. Conversely, a lack of feedback eroded belief and discouraged players from engaging with the reporting process — if a player did not get feedback, they may think their report was being ignored, e.g. *"if I report somebody and I don't get any notification, I feel like they don't take it seriously"* (P18). How feedback was presented mattered as well:

> *"It would be nice to know the actual punishment that was given to the person. Right now, you just get a message, but it doesn't tell you what the thing was"* - P2

This missing information raised doubt within this participant about whether their reports were truly being processed or if the feedback was merely superficial, as P2 continued: *"I feel like it's a very placebo type of thing that they put in"*.

Transparency of the Entire Reporting and Punishment Process. The desire for feedback emerged directly from the lack of transparency often associated with reporting systems, including ambiguity around how reports are handled, the criteria for determining punishments, and inconsistencies in the outcomes of reported cases, which could reduce trust that their reporting matters and mediate the expectancy of their reports. P4 expressed that this lack of transparency led them to suspect that reporting could be influenced by personal bias. They shared that, *"the decision-making process can be unclear. Punishments might vary, which is why I don't fully trust the reporting system"* (P4). Similarly, P10 cynically suggested that this opacity and vagueness might be an intentional design by the developers:

> *"I definitely think it's subjective, and they kind of make it vague on purpose so that it's easier for them to pick and choose. This way, they can defend whatever decisions they make because the rules are intentionally unclear"* - P10

The lack of transparency could also tie into the heavy automation present in many moderation systems for large multiplayer games[2]. Players held varying levels of trust in reports processed by an AI versus by a human, which would impact how they expected their report to be handled. Overall, the general trend was an increased trust in human moderation, as players believed that fellow humans were better equipped to handle complex cases and ambiguous situations requiring emotional judgment:

> *"The advantage of humans handling the report system is their ability to understand the context of a situation, including factors like game state, player intent, and chat history* - P4

However, some players argued that AI may perform better than humans in certain scenarios because, *"every human can make errors"* (P3). Players also acknowledged AI's ability to handle reports on a large scale efficiently, as P18 highlighted, *"Every day, I'm sure there are more than 100,000 reports. Many people are gaming, and it'd be difficult for a team of human staff to review every single one of them"*. Some players appreciated that AI systems can act as an early warning mechanism for inappropriate behaviour. P15 stated, *"if bad content or a violation of the community guidelines is coming from your account, you should receive a warning for it."*

4.4 RQ4: How do the prior factors play a role in broader perceptions of trust and expectations regarding the game?

This section considers how the factors that affect player expectancy translate into overarching attitudes about the game but its surrounding ecosystem, including its players, the developer's reputation, and the overall experience regarding toxicity and perceived fairness.

4.4.1 Attitudes Towards Game Developers and the Broader Ecosystem. While attitudes towards game developers can affect players' expectancy for outcomes, the opposite also holds — the a priori factors that affect expectancy can also affect future trust and attitudes, namely, how players perceive the role of the developer and their reputation regarding accepting or rejecting toxic behaviour:

> *"I feel like Riot [author note: Riot = Riot Games, a game developer and publisher company notable for League of Legends and Valorant] expects a level of toxicity that they deem as like coming with this type of competitive game... maybe when I play this game I should go in expecting like a baseline level of toxicity"* - P6

Participants admitted the fact that game companies are likely fine with having some toxic players on their platform as well, because *"taking them off their platform would lose money"* (P16) and *"if there's more players ... they're gonna have income"* (P15). Yet, although players highlight these issues, their continued gameplay suggests a complex relationship between frustration with toxicity in the game, cynicism with developer motives, reporting habits, but also a desire to continue playing.

In contrast, positive factors affecting expectancy in the reporting system can facilitate goodwill towards the game developer and encourage continued gameplay. For P19 (who plays *Valorant*), the sentiment *"that action is being taken... I still trust that the [reporting] process is done appropriately"* is important, as it affects *"how much I would enjoy the game"*. P19 mentions that *Valorant* is:

> *"one of the few games where girls also play... I have girl friends who share the same sentiment, where they also feel the need to report players when they say unacceptable things"* - P19

---

[2]e.g. https://support-leagueoflegends.riotgames.com/hc/en-us/articles/201752884-Player-Reporting-Guide-and-FAQ

Overall, our findings here suggest that reporting and punishment systems not only impact in-game experience but also play a significant role in shaping the broader ecosystem around specific games, including highlighting the motives of the developer and promoting positive outcomes such as diversity and inclusivity.

4.4.2   Normalization and Acceptance of Toxic Behaviour. Perspectives on how reporting and moderation are handled led to questions on whether toxic behaviour can ever be truly addressed in games. From a more cynical approach, P7 indicated that perceived ineffectiveness of the reporting system normalizes toxicity, because *"everyone's coming on to play a game to sort of chill and blow off some steam... maybe just being toxic or something helps them"* while knowing they can get away with it. Many participants indicated how toxicity was nowadays inherently ingrained in certain games:

> *"[League of Legends] is a game that's so old that people kind of know what to expect from it"* - P2

> *"I feel like [toxicity]'s just kind of expected, especially with let's say Call of Duty"* - P9

With an ingrained reputation and the perceived ineffectiveness of the primary tool players can use to combat toxicity, players indicated a sense of helplessness — *"it's kind of impossible [to change] because just like how gaming in general inherently is"* (P15) — even when reporting is not the sole way of addressing toxicity. Despite this, when people give up on thinking that the community can ever improve, they lose one of the two motivators for reporting. Some participants felt discouraged about their actions when they felt nothing was changing. Yet when reporting remains the primary in-game option to combat toxicity, this leads to a potentially negative feedback loop.

On the other hand, not everyone felt hopeless, with some participants indicating that the gaming community was getting better. These observations helped participants believe that things were changing positively, and partially played a role in their continued reporting behaviours even if community improvement was not directly solely correlated to reporting behaviours, such as:

> *"going back to like Xbox Call of Duty lobbies ... those were just a cesspool compared to today; there definitely are stronger moderation standards"* - P14

Thus, if players perceived the community as getting better (whether due to reporting practices or not), then their altruistic motivations for reporting are met and they may be more motivated to continue to report (e.g. *"I feel like I'm reporting in that game, it makes the environment better for other players as well. So that gives me more motivation to report in that game"* — P19), leading to a positive feedback loop in terms of change.

5   Discussion

Our findings highlight that, under an EVT perspective, reporting holds two somewhat contrasting sources of value, one driven through retributive motivations to punish the toxic perpetrator, and one driven through altruistic motivations to improve the gaming community. However, participants noted that reporting neither consistently nor effectively meets each of these outcomes, because not every instance of reporting leads to punishment, and punishment is not always conducive to the change required for broader improvement.

We highlight that the belief that such outcomes will be reached (i.e., expectancy) is shaped by various factors such as trust based on prior experiences, normative alignment with the community, and system transparency. Importantly, belief in reporting contexts is shaped, not by self-efficacy like prior EVT studies [63, 74], but by surrounding sociotechnical factors. Participants sometimes indicated low expectancy when they felt like the developer did not care, or the reporting system was too opaque, even when reporting was valued. In our discussion, we expand on two problems related to each of the EVT constructs. For expectancy, we consider ways for developers to increase

user belief in the system and their reporting actions; for value, we consider ways to align two contrasting motivations that may sometimes have misaligned consequences.

5.1 Bridging the Gap of Belief — Design and Challenges Regarding Expectancy

Various factors that mediate people's **beliefs** affect perceptions of reporting's effectiveness. Much prior past work highlights the doubt that players can hold towards perceived ineffective reporting systems [36, 60]. As with our work, players ascribe part of that fault to the developers of the game, and developers play a role in controlling these behaviours [26, 41]. Our work highlights how various dimensions, such as transparency, feedback provision, and prior experiences, all play a role in establishing people's trust in the black-box reporting system and, thus, expectancy.

In Section 2.2, we reviewed how laws that align with shared values encourage adherence, but a lack of transparency in the legal system can influence individuals' likelihood of breaking laws [79]. Analogizing to a gaming context, players emphasized the need for greater transparency in the reporting system and called for developers to prioritize and expedite the handling of reports. A lack of transparency and consistency in regulation lowered their trust in the system, as players could seemingly get away with infractions. Beyond regulation transparency, the provision of feedback was noted as helpful for people to know that their reports are being heard, as well as to build player trust in the system; agreeing with prior work [33]. Given the importance of transparency and feedback, we propose that existing game reporting systems could better align with the existing Santa Clara Principles [70] by being more open about the moderation process and providing detailed explanations to both the reporter and the punished player.

5.1.1 Moderation Trust with Automation. Developers often face challenges in monitoring games and handling reports due to the large number of players. While Hilbers suggests incorporating automation to guide player behaviour [26], we find that players perceived automation as struggling to comprehend complex contexts and deliver fair decisions. Human moderators, on the other hand, were perceived to provide deeper contextual understanding but were limited by resource constraints. Human judgment and bias are tied to player concerns about the inconsistency of reporting outcomes over time, emphasizing the need for a more robust system. This generally aligns with Sparrow et al.'s prior research, which highlights the many roles and mental models of AI moderation in multiplayer games — as an unreliable police force, an unscrupulous governor, an uncaring judge, and an untiring assistant [75].

The problem, common to both Sparrow et al.'s work and our findings, is that players have inconsistent mental models when faced with a black-box moderation system. Yet, we note that while AI moderation adds to this opacity, reporting systems would likely still have a level of higher judgment invisible to the users; even crowdsourced judgments had concerns over transparency [37]. When faced with an uncertain system and uncertain consequences, players inevitably face challenges in trust [76]. Given that trust then delegates to what players *can* observe, we propose rethinking and redesigning human-AI systems for moderation [40] with a focus on communication, transparency, and player perception. We consider systems that might leverage AI to support human decision-making at scale, combining the efficiency of AI with the nuanced judgment of human moderators to enhance fairness and consistency in the reporting process [46, 82].

We highlight ways that AI moderation can be designed to be more acceptable and trustworthy by players, with the assumption of benevolent intention to increase trust [43]. To provide feedback desired by reporters, we envision an AI-supported triage system. AI could escalate the highest priority reports [52] or cluster similar reports [68]. Overall, to extend transparency in driving trust, reporters could view the status of their report at all times, from when it is in this AI-supported queue, to when it is being reviewed by humans or AI, to whether punishment is levied or not and

why, and perhaps even afterwards, whether the offending player has actually reformed. Under this perspective, AI would serve as a proxy for information that is presently missing for reporters.

Instead of optimizing and presenting the AI (and moderation as a whole) as an authoritarian source of truth, developers should highlight and communicate how they are using the AI in their moderation pipeline, provide examples of AI judgment, and provide documentation and statistics for transparency purposes [43]. To garner trust, AI-assisted processes should acknowledge that AI can make mistakes [93]; in such cases, perhaps opportunities to appeal are warranted [75]. In general, trustworthiness in this context is fostered through participants forming a mental model in which they understand, are confident in, respect, and are comfortable with automation [71]. If developers are to use automation as part of their report-handling, then ultimately trust, and thus user expectancy, can only be built through understanding the system rather than fearing it.

5.1.2 Trade-offs for Transparency. Yet, while transparency is vital for trust [7, 43, 93], we acknowledge that complete transparency might not be appropriate. Specifically, revealing the full details of how the reporting system detects and reviews toxic behaviour could enable toxic players to manipulate the system and avoid detection. This issue parallels the broader debate about the extent to which laws should list reasons for crimes. Some scholars argue that listing reasons helps clarify legal boundaries [25] and reduces misuse of the law [24], thereby preventing unintentional violations and protecting citizens' rights. However, others caution that overly detailed reasoning can introduce significant drawbacks such as increased legal complexity [72] and create exploitable loopholes [61, 77]. In a reporting context, if users know a word would definitely be flagged by the rules, then they may be able to bypass it by using a variant (e.g. by changing a letter). Thus, a potential compromise is adopting principle-based provisions, where general guidelines are established, and specific applications are clarified through judicial interpretations or guiding cases [3, 45].

Presently, community guidelines are usually provided upfront *once* but not reviewed as an ongoing activity. We propose weaving community guidelines into the constant gameplay experience where possible, for instance, rotating important guidelines on waiting screens or log-in pages with specific examples or interpretations, or perhaps even incorporating them during periods of low-action into the gameplay loop (e.g. as reminders in the messaging system). This would serve several purposes in improving the expectancy of reporting outcome — highlight that toxicity is a problem that is taken seriously within the game, raise ongoing community awareness of toxicity (and that reporting is a tool to combat it), and reinforce community standards on toxicity that can otherwise be subjective. We also highlight that there are ways for developers to demonstrate transparency beyond revealing details of the process, such as providing broader metrics, diagnoses, stats, and documentation regarding moderation [43]. Providing such broader, anonymized reports may help with transparency and accountability without the ethical concerns (of revealing details to be taken advantage of) or privacy concerns (of naming and shaming, which could lead to witch hunts [6]).

5.2 Building a Better Community - Understanding and Aligning Reporting Values

Our findings suggest that reporting was expected to lead to punishment, and while punishment was regarded as an effective way to reduce toxic behaviour (although possibly bypassed through new accounts), reporting was not always perceived to be effective in reaching the punishment stage [36, 60]. Kou also found that reporting is used for reasons outside of dealing with toxicity, such as its use in coordinated, targeted contexts or its use in venting frustration, dissociating reporting from actual "toxicity" [36]. We find that reporting is used for two sometimes conflicting motivations — as an idealized way to combat toxicity, and as a self-fulfilling way to punish people whom you have had a poor experience with. Yet, despite our participants' distrust of reporting systems that agree with prior research, almost all participants continued to use it (also refer to Q9 in Table 3) in

hopes of a better community, representing a possibly Sisyphean hope that reporting can be used in its ideal context as leading to a better community.

5.2.1 Moving from Individual Punishment to Broader Community Improvement. Beres et al. suggested that an impediment to reporting is the normalization of toxicity — people lack the motivation to report when toxic behaviours are rationalized within the community context or when they distance themselves from the other players [5]. The latter concept of "distancing" was also reinforced through Kou's consideration of permanent bans [35]. Whereas permanent bans are used as a disciplinary method to induce change, Kou highlights how their deeper functions are to produce a stereotype of "the most toxic" player in the community [35]. These prior works highlight the deep division of the community into "toxic" and "non-toxic" players as two groups in tension; this division both motivates and undermines the act of reporting. On one hand, it supports reporting as a way to exclude toxic players (the retributive motivation); on the other, it fosters skepticism about its effectiveness, as the community may not always improve (the altruistic motivation). Poeller and Robinson highlight how reporting fails to address the underlying problem regarding toxic communities [59]. The fracture in the community contrasts with more recently suggested restorative justice approaches to moderation. Restorative justice defines infractions as a violation of people and punishment, rather than a violation of the rules; justice comes from repair and reconciliation rather than punishment [14, 80, 83]. Kannabiran espouses the importance of forgiveness when human relationships are transgressed, with repair coming from a chance to make amends and admit mistakes [31]. Vasalou et al. implemented a system for repairing trust based on apology and forgiveness, finding that these factors allowed victims to recover trust [81].

Restorative justice and forgiveness are promising directions for the altruistic goals of reporting. Yet there are challenges, as we highlight three interrelated open questions. Firstly, Xiao et al. indicated that restorative justice is a community effort, yet victims, offenders, and moderators in gaming communities may lack a shared sense of community [89]; — thus, beyond reporting, *how can we reinforce the shared identity of players to re-form a community as a whole, especially given the ephemeral context of team-based multiplayer games*? Secondly, Kannabiran brought up challenges with restorative justice and forgiveness as a whole [31] — thus, *how do we ensure that forgiveness takes place after reporting in a manner that is adequate, appropriate, and fair to the victim*? Lastly, Reid et al. explored novel ways for people to deal with toxicity, but also found that people found that players were skeptical of tools that focus on support rather than discipline [65]. People desire discipline as a form of revenge against toxicity — thus, *how do we validate the negative experience and deal with the frustration and spite that a reporting player might have held towards the other*?

5.2.2 Finding Common Values. These challenges converge on the issue of **aligning the sometimes divergent reporting motivations towards a unified goal**. Punishment and restorative justice are not inherently oppositional, but punishment must be *meaningful* to induce change. Our challenges double as design suggestions. For instance, we consider solutions such as fostering a shared community identity through value alignment. Doing so fundamentally may require re-thinking design from the ground up, to reconcile opposing values between reporting players and reported players, and even within groups of reporting players. Design methodologies such as Value-Sensitive Design [56] and Ally-Opponent Understanding [29] provide promising baselines for building with multiple groups of values. Both these approaches espouse the importance of talking to stakeholders early, understanding their perspectives, and evaluating by multiple groups of stakeholders.

Although we have not engaged in such a design exercise, we provide some initial brainstormed examples. Games could provide non-identifying glimpses and light tasks about the other players to humanize them and support camaraderie in initial team-building (e.g. "this player has not played in several months — help them adapt!"). In doing so, we point towards a shared goal that can be

accomplished collaboratively, even in a competitive environment. We also encourage restorative interactions, such as an apology by the toxic player, and empowering the victim of toxic behaviour to articulate the impact of toxicity and involve them in the resolution. These steps may help bridge the gap between punitive and altruistic approaches, advancing toward a better community.

6 Limitations and Future Work

Limitations are mainly tied to the participants and their engagement in the study. Some of these limitations are fundamental to user studies, for example, self-selection bias [20] — people who are likely to want to contribute to the study may already have some vested opinions in reporting systems compared to those who do not, which may skew the presented perspectives. As our research focuses specifically on reporting practices, social desirability response bias may also exist — people may try to respond to paint themselves in a morally correct light [64]. In this study, we hypothesize that such effects may cause people to highlight the altruistic motivation for reporting and minimize the retributive motivations. This could be exacerbated by the fact that our study focused on people's recounting of their experiences, which may be subject to recall bias due to selective memory [73]. Reporting could take place as singular moments far in the past, meaning that recollective feelings may have changed or become muted over time. To alleviate these problems, future work can incorporate a general sample of gameplay to verify whether the recounted data holds during actual observed experiences across a wider sample of players.

We also highlight that our work is still exploratory in understanding people's reporting systems and their relationship to EVT. While our qualitative insights provide an initial understanding of our participants' values and expectancies, future work can perform a confirmatory survey with a large sample of participants using existing scales for motivation and self-efficacy (e.g. as in [74, 90]), to see if similar results hold through further triangulation [13].

Extending on the sample of interview participants, most were young. Although this was somewhat expected given the domain and samples seen in prior research, participants did mention that age can affect players' toxic behaviour and their engagement with the reporting system, for example, several participants mentioned that they had been more likely to abuse the report button when they were younger. Most participants also discussed competitive or team-based games; future work could explore reporting experiences in co-operative games. Finally, the general opacity of different reporting systems makes it hard to make concrete judgments; many perspectives in this work rely on participant anecdotes, experiences, and their best guesses.

7 Conclusion

We holistically investigated reporting systems in multiplayer online video games through expectancy-value theory. Using a survey and follow-up interviews, we found that players' values in reporting are both altruistic and retributive, viewing reporting as a means of community betterment and personal venting, respectively. However, players often perceive that their desired outcomes expected from these motivations are not always fully realized, with the degree to which players believe in these outcomes influenced by factors such as the players' personal attitudes, their trust in the community and the developer, and design factors of the reporting system including its level of transparency, the provision of feedback, and the degree of automation. These perceptions affect players' beliefs and attitudes, shaping their view of developer reputation and potentially leading to the normalization or passive acceptance of toxic behaviour in games. Based on our findings, we propose guidelines for designing reporting systems and fostering healthier gaming communities.

References

[1] Leigh Achterbosch, Charlynn Miller, and Peter Vamplew. 2017. A Taxonomy of Griefer Type by Motivation in Massively Multiplayer Online Role-Playing Games. *Behaviour & Information Technology* 36, 8 (Aug. 2017), 846–860. https://doi.org/10.1080/0144929X.2017.1306109

[2] Leigh Achterbosch, Peter Vamplew, and Evita March. 2024. Assessing the Impact of Griefing in MMORPGs Using Self-Determination Theory. *Computers in Human Behavior* 161 (Dec. 2024), 108388. https://doi.org/10.1016/j.chb.2024.108388

[3] Robert Alexy. 2010. *A theory of constitutional rights.* Oxford university press.

[4] George Beknazar-Yuzbashev, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346* (2022).

[5] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445157

[6] Kelly Bergstrom. 2011. "Don't Feed the Troll": Shutting down Debate about Community Expectations on Reddit.Com. *First Monday* (July 2011). https://doi.org/10.5210/fm.v16i8.3498

[7] Shailee Bhatia, Vatsal Mittal, and Saransh Sabharwal. 2024. Building Trust and Transparency in AI: A Review of Explainable AI and Its Ethical Implications. In *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*. 650–655. https://doi.org/10.1109/EmergIN63207.2024.10961081

[8] Stephanos Bibas. 2006. Transparency and participation in criminal procedure. *NYUL rev.* 81 (2006), 911.

[9] Jeremy Blackburn and Haewoon Kwak. 2014. STFU NOOB! Predicting Crowdsourced Decisions on Toxic Behavior in Online Games. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. Association for Computing Machinery, New York, NY, USA, 877–888. https://doi.org/10.1145/2566486.2567987

[10] Virginia Braun and Victoria Clarke. 2021. *Thematic analysis: A Practical Guide.* Sage Publications, Thousand Oaks, California, United States, 57–71.

[11] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[12] Alessandro Canossa, Dmitry Salimov, Ahmad Azadvar, Casper Harteveld, and Georgios Yannakakis. 2021. For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY (Oct. 2021), 253:1–253:29. https://doi.org/10.1145/3474680

[13] Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J. Neville. 2014. The Use of Triangulation in Qualitative Research. *Oncology Nursing Forum* 41, 5 (Sept. 2014), 545–547. https://doi.org/10.1188/14.ONF.545-547

[14] Janine Natalya Clark. 2008. The Three Rs: Retributive Justice, Restorative Justice, and Reconciliation. *Contemporary Justice Review* 11, 4 (Dec. 2008), 331–350. https://doi.org/10.1080/10282580802482603

[15] Russell Cropanzano, Marion Fortin, and Jessica F Kirk. 2015. How do we know when we are treated fairly? Justice rules and fairness judgments. In *Research in personnel and human resources management*. Vol. 33. Emerald Group Publishing Limited, 279–350.

[16] John M Darley. 2009. Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science* 5, 1 (2009), 1–23.

[17] Saraswati Dawadi, Sagun Shrestha, and Ram A. Giri. 2021. Mixed-Methods Research: A Discussion on Its Types, Challenges, and Criticisms. *Journal of Practical Studies in Education* 2, 2 (Feb. 2021), 25–36. https://doi.org/10.46809/jpse.v2i2.20

[18] Edward L. Deci, Anja H. Olafsen, and Richard M. Ryan. 2017. Self-Determination Theory in Work Organizations: The State of a Science. *Annual Review of Organizational Psychology and Organizational Behavior* 4, Volume 4, 2017 (March 2017), 19–43. https://doi.org/10.1146/annurev-orgpsych-032516-113108

[19] Jacquelynne S. Eccles and Allan Wigfield. 2020. From Expectancy-Value Theory to Situated Expectancy-Value Theory: A Developmental, Social Cognitive, and Sociocultural Perspective on Motivation. *Contemporary Educational Psychology* 61 (April 2020), 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

[20] Dirk M. Elston. 2021. Participation Bias, Self-Selection Bias, and Response Bias. *Journal of the American Academy of Dermatology* 0, 0 (June 2021). https://doi.org/10.1016/j.jaad.2021.06.025

[21] David S Evans. 2012. Governing bad behavior by users of multi-sided platforms. *Berkeley Tech. LJ* 27 (2012), 1201.

[22] Jessica Kay Flake, Kenneth E. Barron, Christopher Hulleman, Betsy D. McCoach, and Megan E. Welsh. 2015. Measuring Cost: The Forgotten Component of Expectancy-Value Theory. *Contemporary Educational Psychology* 41 (April 2015), 232–244. https://doi.org/10.1016/j.cedpsych.2015.03.002

[23] Julian Frommel and Regan L Mandryk. 2024. 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)*. Association for

Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3613904.3642523

[24] Lon Luvois Fuller. 1969. The morality of law. (1969).

[25] Herbert Lionel Adolphus Hart and Leslie Green. 2012. *The concept of law.* oxford university press.

[26] J. J. P. Hilbers. 2015. *An Evaluation of the Effectiveness of Smite's Player Report System.* Master's thesis.

[27] Kristine Ashley Holmes. 2024. *Toxicity in the Online Game League of Legends.* Master's thesis. San Diego State University.

[28] Jonathan Jackson, Ben Bradford, Mike Hough, Andy Myhill, Paul Quinton, and Tom R Tyler. 2012. Why do people comply with the law? Legitimacy and the influence of legal institutions. *British journal of criminology* 52, 6 (2012), 1051–1071.

[29] Sukwoo Jang, Minsun Hong, and Youn-kyung Lim. 2019. Ally-Opponent Understanding: Co-existence of Conflicting Values through Participatory Design. *Design Works* 2, 2 (2019), 16–28.

[30] Justin Juvrud. 2020. *"ROFL F*ck You" : Understanding the Current State of Toxicity in Battlefield V.*

[31] Gopinaath Kannabiran. 2021. I Am Sorry! *interactions* 28, 5 (Aug. 2021), 14–15. https://doi.org/10.1145/3477089

[32] Jawad Khan, Imran Saeed, Muhammad Zada, Amna Ali, Nicolás Contreras-Barraza, Guido Salazar-Sepúlveda, and Alejandro Vega-Muñoz. 2022. Examining whistleblowing intention: The influence of rationalization on wrongdoing and threat of retaliation. *International journal of environmental research and public health* 19, 3 (2022), 1752.

[33] Bastian Kordyaka and Björn Kruse. 2021. Curing Toxicity – Developing Design Principles to Buffer Toxic Behaviour in Massive Multiplayer Online Games. *Safer Communities* 20, 3 (Jan. 2021), 133–149. https://doi.org/10.1108/SC-10-2020-0037

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

[35] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 334:1–334:21. https://doi.org/10.1145/3476075

[36] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445279

[37] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 62:1–62:17. https://doi.org/10.1145/3134697

[38] Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* 120, 7 (2023), e2210666120.

[39] Samuli Laato, Bastian Kordyaka, and Juho Hamari. 2024. Traumatizing or Just Annoying? Unveiling the Spectrum of Gamer Toxicity in the StarCraft II Community. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642137

[40] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–18.

[41] Matthew Lapolla. 2020. Tackling Toxicity: Identifying and Addressing Toxic Behavior in Online Video Games. *Seton Hall University Dissertations and Theses (ETDs)* (May 2020).

[42] William Laumann. 2021. *'You cannot mute that someone walks under a turret and dies':: Exploring League of Legends-players' perception, definitions of toxicity and their effect on everyday life.* Master's thesis. Uppsala University.

[43] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).* Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[44] Chwen-Yea Lin, Wei-Hsi Hung, Kwoting Fang, and Chien-Chung Tu. 2015-10-05Z. Understanding Players' Achievement Values from MMORPGs: An Exploratory Study. *Internet Research* 25, 5 (2015-10-05Z), 829–851. https://doi.org/10.1108/IntR-12-2013-0268

[45] Karl N Llewellyn. 2012. *The bramble bush: On our law and its study.* Quid Pro Books.

[46] Thodoris Lykouris and Wentao Weng. 2024. Learning to defer in content moderation: The human-ai interplay. *arXiv preprint arXiv:2402.12237* (2024).

[47] Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581097

[48] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 278:1–278:30. https://doi.org/10.1145/3610069

[49] Kirsti Malterud, Volkert Dirk Siersma, and Ann Dorrit Guassora. 2016. Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research* 26, 13 (Nov. 2016), 1753–1760. https://doi.org/10.1177/1049732315617444

[50] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity Detection in Multiplayer Online Games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*. 1–6. https://doi.org/10.1109/NetGames.2015.7382991

[51] Brian McInnis, Leah Ajmani, Lu Sun, Yiwen Hou, Ziwen Zeng, and Steven P Dow. 2021. Reporting the Community Beat: Practices for Moderating Online Discussion at a News Website. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[52] Lloyd Montgomery, Daniela Damian, Tyson Bulmer, and Shaikh Quader. 2018. Customer Support Ticket Escalation Prediction Using Feature Engineering. *Requirements Engineering* 23, 3 (Sept. 2018), 333–355. https://doi.org/10.1007/s00766-018-0292-3

[53] Katherine Muenks, J. E. Miller, Brendan A. Schuetze, and Tiffany A. Whittaker. 2023. Is Cost Separate from or Part of Subjective Task Value? An Empirical Examination of Expectancy-Value versus Expectancy-Value-Cost Perspectives. *Contemporary Educational Psychology* 72 (Jan. 2023), 102149. https://doi.org/10.1016/j.cedpsych.2023.102149

[54] Benjamin Nagengast, Herbert W. Marsh, L. Francesca Scalas, Man K. Xu, Kit-Tai Hau, and Ulrich Trautwein. 2011. Who Took the "×" out of Expectancy-Value Theory?: A Psychological Mystery, a Substantive-Methodological Synergy, and a Cross-National Generalization. *Psychological Science* 22, 8 (Aug. 2011), 1058–1066. https://doi.org/10.1177/0956797611415540

[55] Louise Anker Nexø and Søren Kristiansen. 2023. Players Don't Die, They Respawn: A Situational Analysis of Toxic Encounters Arising from Death Events in League of Legends. *European Journal on Criminal Policy and Research* 29, 3 (Sept. 2023), 457–476. https://doi.org/10.1007/s10610-023-09552-y

[56] Ilse Oosterlaken. 2015. Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An Exploration. *Science and Engineering Ethics* 21, 2 (April 2015), 359–379. https://doi.org/10.1007/s11948-014-9536-x

[57] Leon Petrazycki and A Javier Trevino. 2017. *Law and morality.* Routledge.

[58] Susanne Poeller, Martin Johannes Dechant, Madison Klarkowski, and Regan L. Mandryk. 2023. Suspecting Sarcasm: How League of Legends Players Dismiss Positive Communication in Toxic Environments. *Proc. ACM Hum.-Comput. Interact.* 7, CHI PLAY (Oct. 2023), 374:1–374:26. https://doi.org/10.1145/3611020

[59] Susanne Poeller and Raquel B. Robinson. 2024. 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)*. Association for Computing Machinery, New York, NY, USA, 282–284. https://doi.org/10.1145/3665463.3678863

[60] Antti Pohjanen. 2018. Report Please! : A Survey on Players' Perceptions towards the Tools for Fighting Toxic Behavior in Competitive Online Multiplayer Video Games. (2018).

[61] Richard A Posner. [n. d.]. ECONOMIC ANALYSIS OF LAW.

[62] David W. Putwain, Laura J. Nicholson, Reinhard Pekrun, Sandra Becker, and Wendy Symes. 2019. Expectancy of Success, Attainment Value, Engagement, and Achievement: A Moderated Mediation Analysis. *Learning and Instruction* 60 (April 2019), 117–125. https://doi.org/10.1016/j.learninstruc.2018.11.005

[63] Arif Rachmatullah, Frieda Reichsman, Trudi Lord, Chad Dorsey, Bradford Mott, James Lester, and Eric Wiebe. 2021. Modeling Secondary Students' Genetics Learning in a Game-Based Environment: Integrating the Expectancy-Value Theory of Achievement Motivation and Flow Theory. *Journal of Science Education and Technology* 30, 4 (Aug. 2021), 511–528. https://doi.org/10.1007/s10956-020-09896-8

[64] Donna M. Randall and Maria F. Fernandes. 1991. The Social Desirability Response Bias in Ethics Research. *Journal of Business Ethics* 10, 11 (Nov. 1991), 805–817. https://doi.org/10.1007/BF00383696

[65] Elizabeth Reid, Regan L. Mandryk, Nicole A. Beres, Madison Klarkowski, and Julian Frommel. 2022. Feeling Good and In Control: In-game Tools to Support Targets of Toxicity. *Proc. ACM Hum.-Comput. Interact.* 6, CHI PLAY (Oct. 2022), 235:1–235:27. https://doi.org/10.1145/3549498

[66] Sarah T Roberts. 2017. *Content moderation.*

[67] Isabela Maria Rosal Santos, Felipe Silva, José Renato Laranjeira de Pereira, Mariane Andrade Moreira, Paulo Henrique Atta Sarmento, and Thiago Guimarães Moraes. 2020. Contribution for Santa Clara Principles-A contribution regarding transparency around the use of automated tools and decision-making. (2020).

[68] Suman Roy, Durga Prasad Muni, John-John Yeung Tack Yan, Navin Budhiraja, and Fabien Ceiler. 2016. Clustering and Labeling IT Maintenance Tickets. In *Service-Oriented Computing*, Quan Z. Sheng, Eleni Stroulia, Samir Tata, and Sami Bhiri (Eds.). Springer International Publishing, Cham, 829–845. https://doi.org/10.1007/978-3-319-46295-0_58

[69] Richard M. Ryan, C. Scott Rigby, and Andrew Przybylski. 2006. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30, 4 (Dec. 2006), 344–360. https://doi.org/10.1007/s11031-006-9051-8

[70] Santa Clara Principles Working Group. 2018. The Santa Clara Principles on Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org Accessed: 2025-01-26.

[71] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors* 58, 3 (May 2016), 377–400. https://doi.org/10.1177/0018720816634228

[72] Frederick Schauer. 1991. *Playing by the rules: A philosophical examination of rule-based decision-making in law and in life.* Clarendon Press.

[73] Alan L. Shields, Saul Shiffman, and Arthur Stone. 2010. Recall Bias: Understanding and Reducing Bias in PRO Data Collection. In *ePro.* Routledge.

[74] Huiyoung Shin and Sunjeong Gyeong. 2025. When Motivation Isn't Enough: Perceived Teacher Attitudes, Efficacy, and Efforts Moderate Student Responses to Bullying. *BMC Psychology* 13, 1 (June 2025), 609. https://doi.org/10.1186/s40359-025-02959-8

[75] Lucy A. Sparrow, Ren Galwey, Dahlia Jovic, Taylor Hardwick, and Mahli-Ann Butt. 2024. Towards Ethical AI Moderation in Multiplayer Games. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (Oct. 2024), 344:1–344:30. https://doi.org/10.1145/3677109

[76] Robin Steedman, Helen Kennedy, and Rhianne Jones. 2020. Complex Ecologies of Trust in Data Practices and Data-Driven Systems. *Information, Communication & Society* 23, 6 (May 2020), 817–832. https://doi.org/10.1080/1369118X.2020.1748090

[77] Cass R Sunstein. 2018. *Legal reasoning and political conflict.* Oxford University Press.

[78] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376191

[79] Tom R Tyler. 2001. Obeying the law in America: procedural justice and the sense of fairness. *Issues of Democracy* 6, 1 (2001), 16–21.

[80] Mark S. Umbreit. 1989. Crime Victims Seeking Fairness, Not Revenge: Toward Restorative Justice. *Federal Probation* 53 (1989), 52.

[81] Asimina Vasalou, Astrid Hopfensitz, and Jeremy V. Pitt. 2008. In Praise of Forgiveness: Ways for Repairing Trust Breakdowns in One-off Online Interactions. *International Journal of Human-Computer Studies* 66, 6 (June 2008), 466–480. https://doi.org/10.1016/j.ijhcs.2008.02.001

[82] Leijie Wang and Haiyi Zhu. 2022. How are ml-based online content moderation systems actually used? studying community size, local activity, and disparate treatment. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 824–838.

[83] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. 2008. Retributive and Restorative Justice. *Law and Human Behavior* 32, 5 (2008), 375–389. https://doi.org/10.1007/s10979-007-9116-6

[84] Allan Wigfield. 1994. Expectancy-Value Theory of Achievement Motivation: A Developmental Perspective. *Educational Psychology Review* 6, 1 (March 1994), 49–78. https://doi.org/10.1007/BF02209024

[85] Allan Wigfield and Jacquelynne S. Eccles. 2000. Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 68–81. https://doi.org/10.1006/ceps.1999.1015

[86] Allan Wigfield and Jessica R. Gladstone. 2019. What Does Expectancy-value Theory Have to Say about Motivation and Achievement in Times of Change and Uncertainty? In *Motivation in Education at a Time of Global Change.* Vol. 20. Emerald Publishing Limited, 15–32. https://doi.org/10.1108/S0749-742320190000020002

[87] Michel Wijkstra, Katja Rogers, Regan L. Mandryk, Remco C. Veltkamp, and Julian Frommel. 2023. 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).* Association for Computing Machinery, New York, NY, USA, 3–9. https://doi.org/10.1145/3573382.3616068

[88] Michel Wijkstra, Katja Rogers, Regan L. Mandryk, Remco C. Veltkamp, and Julian Frommel. 2024. 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), 315:1–315:32. https://doi.org/10.1145/3677080

[89] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6 (Sept. 2023), 83:1–83:36. https://doi.org/10.1145/3603625

[90] Shu Ching Yang, Chiao-Ling Huang, and An-Sing Chen. 2013. An Investigation of College Students' Perceptions of Academic Dishonesty, Reasons for Dishonesty, Achievement Goals, and Willingness to Report Dishonest Behavior. *Ethics & Behavior* 23, 6 (Nov. 2013), 501–522. https://doi.org/10.1080/10508422.2013.802651

[91] Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2024. Game On, Hate Off: A Study of Toxicity in Online Multiplayer Environments. *ACM Games* 2, 2 (Aug. 2024), 14:1–14:13. https://doi.org/10.1145/3675805

[92] Ceren Yılmaz. 2022. *Understanding gender in PC games and gaming culture: the case of League of Legends.* Master's thesis. Middle East Technical University.

[93] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How Transparency Modulates Trust in Artificial Intelligence. *Patterns* 3, 4 (April 2022). https://doi.org/10.1016/j.patter.2022.100455

[94] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2024. Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Content. https://doi.org/10.48550/arXiv.2309.06688 arXiv:2309.06688 [cs]

[95] Zinan Zhang, Sam Moradzadeh, Andrew Woan, and Yubo Kou. 2024. Toxicity by Game Design: How Players Perceive the Influence of Game Design on Toxicity. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (Oct. 2024), 345:1–345:31. https://doi.org/10.1145/3677110