Exploring Aggressors' InMatch Cognitive and Emotional Formation and Toxic Behavior Trajectories in MOBA Games

Kangyu Yuan
The Hong Kong University of Science and Technology
Hong Kong, China
kyuanaf@connect.ust.hk

Hanfang Lyu
The Hong Kong University of Science and Technology
Hong Kong, China
hanfang.lyu@connect.ust.hk

Runhua Zhang
The Hong Kong University of Science and Technology
Hong Kong, China
runhua.zhang@connect.ust.hk

Hansika Murugu
College of Information
University of Maryland, College Park
College Park, Maryland, USA
hmurugu@umd.edu

Xiaojuan Ma*
The Hong Kong University of Science and Technology
Hong Kong, China
mxj@cse.ust.hk

Abstract

Toxic behavior in Multiplayer Online Battle Arena (MOBA) games has become a major issue. While previous studies have examined factors influencing toxic behavior, few have captured the cognitive and emotional states of the aggressors at the point of emergence of toxic behavior, or traced its evolution across an entire match. To fill the gap, we conducted replay-based semi-structured interviews with 18 players who recently initiated toxic behavior during matches. With adapted retrospective think-aloud protocols and players' emotional journey maps, we collected their subjective perceptions and dynamic changes of emotion. Through thematic analysis, we identified a multi-dimensional criterion for evaluating toxicity severity and a three-layer cognition–emotion association structure, and described recurring persistent and single-instance patterns of toxic behavior observed in our matches. Based on our findings, we contribute to understanding the internal evolution of player toxicity and discuss implications for preventive intervention strategies and designs aiming at mitigating toxic behavior.

CCS Concepts

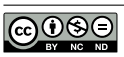
• **Human-centered computing → Computer supported cooperative work**.

Keywords

Toxic behavior, MOBA, Journey map, Retrospective think-aloud

ACM Reference Format:
Kangyu Yuan, Hanfang Lyu, Runhua Zhang, Hansika Murugu, and Xiaojuan Ma. 2026. Exploring Aggressors' InMatch Cognitive and Emotional Formation and Toxic Behavior Trajectories in MOBA Games. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790272

*Corresponding author



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790272

1 Introduction

Content Warning: This manuscript includes explicit instances of toxic language and detailed analysis of toxic behavior.

Multiplayer Online Battle Arena (MOBA) games have emerged as a form of digital entertainment with significant impact, attracting hundreds of millions of players worldwide through highly competitive, collaborative, and fast-paced gameplay [29, 42, 50]. However, player experiences are increasingly threatened by persistent toxic behaviors within games, including, but not limited to, verbal abuse, harassment, and feeding [1, 6, 32, 70]. Such behaviors not only jeopardize players' psychological well-being and social experiences [32, 34], but also undermine the cohesion and sustainability of the game communities, leading to player attrition and churn [20, 49, 52, 70].

To address this issue, both academia and industry have proposed a range of strategies to identify and mitigate toxic behaviors in games, but most of the interventions focus on the victim's side. Some implemented real-time content analysis and monitoring to filter offensive messages before they reach potential victims [23, 51, 69]. Others propose post hoc platform interventions such as reporting, muting, blocking, and account bans [46, 76, 77]. While these mechanisms can mitigate some immediate harm and provide channels for accountability, they also present notable limitations. For the former, their effectiveness may be limited when players deliberately circumvent content filters (*e.g.,* homophones and abbreviations) or employ non-verbal actions (*e.g.,* pings, an in-game reminder function) to perpetuate toxicity. For the latter, victims are often exposed to harm before interventions take effect, and aggressors may perceive sanctions as unjust and respond with frustration rather than behavioral adjustment [21, 41, 48, 79]. A more desirable alternative is to identify opportunities for early intervention to prevent toxicity at its onset, which requires a comprehensive understanding of the aggressor's perspective of toxic behavior evolution within gaming contexts. [76, 77]

Existing research suggested that toxic behaviors are rarely isolated incidents; rather, they often emerge and escalate gradually within complex and dynamic social situations [36, 77]. In the highly

competitive and collaborative environment of MOBA games, the interplay of situational factors such as team performance, role assignments, and group communication can influence players' internal states and behavioral tendencies [36, 49], which can consequently cause toxic actions or suppress escalation. Hence, dissecting the triggers and fluctuations in the game toxic behaviors requires tracking the relationships among events, potential aggressors' cognition, emotions, and actions as a game unfolds.

This paper aims to systematically examine the formation and escalation of toxic behaviors over time by analyzing the experiences, perceptions, and decisions of aggressors during a MOBA game. Specifically, we explored the following three research questions:

- RQ1: What perception do aggressors hold regarding their potential toxic behaviors in games?
- RQ2: What inner states do aggressors experience, and how do they form when initiating toxic behaviors?
- RQ3: What patterns do aggressors' toxic behaviors follow over the course of a match?

However, achieving this goal in real-world gaming environments is challenging, due to the complexity of in-game communication channels, the interplay between internal states and game dynamics, and the constraints on aggressors' self-disclosure [36]. To address these challenges, we conducted a series of qualitative research activities with 18 *League of Legends (LoL)* (a popular MOBA game [53, 68]) players of diverse gaming backgrounds. We first invited each participant to submit a screen recording of a game match. Three researchers reviewed every game recording frame by frame, annotating all potential toxic behaviors initiated or experienced by the participant based on the taxonomy (Appendix A) synthesized from previous work [40, 43, 45, 52]. Next, we adapt the retrospective think-aloud (RTA) method, prompting the participants to detail their thoughts, feelings, and actions while watching their game replay, which was annotated the touchpoints by researchers before. When a pre-annotated event occurred in the video or the participants encountered a game segment with particular personal significance, they were asked to elaborate on the game dynamics, social context, and their nuanced emotional and cognitive experiences. Finally, we engaged the participants in co-creating a player journey map, visualizing their perceived game evolution and the trajectory of emotional changes throughout the match. During the mapping process, they explained the reasons behind these changes and offered direct descriptions of critical moments to facilitate cross-validation with data from RTA.

We applied thematic analysis to the interview transcripts and triangulated different data sources to ensure that our interpretations faithfully reflected the participants' perspectives and experiences. The results reveal that: from the aggressor's perspective, (1) players typically assess the toxicity of their own actions across five dimensions; (2) their behavioral orientation is shaped by perceptions of contextual change, which informs specific action intentions; (3) more negative and intense emotions are generally linked to higher toxicity, while exceptions exist; and (4) their toxic behavior perform a single-instance or persistent pattern.

Despite the limited sample size and specificity to *League of Legends* our study provides a nuanced account of player experiences, offering valuable references for the design of aggressor-oriented early interventions of toxic behaviors in MOBA games. Our contributions to the HCI community are threefold:

- We enrich the understanding of how aggressors in MOBA games evaluate the severity of toxic behaviors by eliciting their own criteria through retrospective think-aloud and journey map.
- We identify recurring patterns in how participants link their internal states to the emergence of toxic behaviors across a MOBA match, using synchronized transcripts, gameplay timelines, and emotional journey maps to show how specific cognition and emotion precede and shape the form of toxicity.
- We translate these empirically observed patterns into game design strategies aimed at reducing the frequency and escalation of toxic behaviors and supporting a healthier gaming ecosystem by aligning intervention timing and feedback with identified triggers and turning points.

2 Related Work

2.1 Attacker-Centered Perspectives on Toxic Behavior in MOBA Games

Toxic behavior in Multiplayer Online Battle Arena (MOBA) games has attracted sustained academic attention due to its high prevalence and substantial negative impact on player experience, community health, and company revenue [20, 21, 32]. Despite this interest, a unified definition of toxic behavior has only emerged recently. Early literature often treated toxicity as an umbrella term for various forms of negative behavior by players in online environments [70], or as disruptive behaviors that are perceived as harmful by others [6]. More recently, Kordyaka et al. [33] proposed a more precise definition, describing toxic behavior as a collective term for actions perceived as disruptive by other players that are not required by gameplay itself. This definition also distinguishes toxic behavior by modality (text, speech, or in-game actions), target (teammates or opponents), intention (internal or external), and timing (proactive or reactive), enabling more systematic identification and categorization in the MOBA context.

Given the importance of mitigating toxicity for designing effective interventions, prior research has examined this phenomenon from multiple perspectives. Empirical game data analyses and literature syntheses describe a behavioral spectrum ranging from minor disturbances to severe abuse [40, 43, 45]. Scholars have also identified various contributing factors, including personality traits, emotional states such as anger, and broader sociocultural orientations [32, 35, 37, 38, 43]. Other studies have explored how players perceive and interpret toxicity, addressing topics such as recognizing toxic communication [58], normalizing toxicity [6], the effect of identity [25], trust and communication dynamics [47], and coping strategies [1].

However, most existing findings stem from researchers' reinterpretations of gameplay data [40], survey-based quantitative studies [25, 32, 36–38, 58], or general perspective interviews and self-reports [6, 35]. Only a few have directly prompted players to recall specific in-game scenarios [1, 47], and these typically focus on communication perceptions or avoidance strategies rather than toxic

behavior itself. This scarcity can be attributed to the unpredictability of in-game toxicity, since not every match involves such incidents, and to privacy concerns that discourage self-reporting [13, 57, 84]. Consequently, first-hand aggressor-side narratives based on real cases remain rare, limiting our understanding of how aggressors interpret their actions and how toxicity unfolds within specific game contexts.

To address this gap, this study focuses on first-person narratives from aggressors, grounded in concrete instances of toxic behavior they exhibited in-game. By examining how aggressors perceive, evaluate, and rationalize their actions, we aim to describe how aggressors in our sample perceive and explain factors they see as driving their toxic behaviors, offering qualitative insights into how toxicity is experienced and narrated within gameplay.

2.2 Fine-grained Exploration on Direct Factors at Toxic Behavior Emergence

Existing research has examined the roots of negative behavior in MOBA games from multiple angles, including individual traits [32, 36, 37], player motivations [7, 35, 42], social dynamics [35, 44, 70], and game mechanisms [8, 83]. While these studies provide valuable theoretical and empirical foundations, they largely emphasize out-of-game factors or isolated incidents and rarely reveal how players' internal states translate into moment-to-moment actions. Little is known about how toxic states are triggered, evolve, and culminate in specific in-game behaviors. This gap in nuanced understanding of real-time pathways limits the precision of intervention strategies, making it difficult to identify proper intervention moments or apply insights effectively in game design [19, 83].

To address this limitation, we apply the General Aggression Model (GAM)[4] to explain how individuals generate aggressive responses in different contexts. GAM posits that aggressive behavior arises from the interaction of cognitive states (*e.g.,* concepts and scripts related to hostility), affective states (*e.g.,* anger, hostility), and physiological arousal, which are integrated through evaluation and decision-making processes to produce aggressive actions. Validated in studies of generalized aggression, this is suitable to the high-pressure, feedback-intensive environment of MOBA games [80]. Our study focuses on its cognitive and emotional dimensions, which are most accessible via self-reports and narratives. Given the dynamic nature of MOBA games, where each instance of communication, tactical adjustment, or competitive shift can alter players' emotions, cognition, and actions [19, 36], a systematic characterization of inner state and behavioral patterns over time is needed.

Based on these needs, our study pursues two main objectives. First, by guiding players to recall specific instances of toxic behavior with a timeline, we aim to capture the cognitive and emotional shifts that occur during moments of toxicity escalation. Second, by mapping a global timeline of in-game events, we seek to identify recurring patterns in how toxic behavior evolves across different stages of play, providing empirical evidence to inform preventive intervention strategies.

2.3 From Reactive to Preventive: A Shift in Toxicity Intervention in MOBA Games

In both commercial practice and academic research on MOBA games, numerous interventions and mechanism optimizations have been developed to address negative behaviors [8, 62, 77]. Mainstream solutions include automated chat moderation and profanity filters, player reporting and muting functions, and post-game penalties such as suspensions, point deductions, or bans [20, 77]. However, most of these approaches act only after malicious behaviors become explicit, relying on post-hoc punishment or real-time blocking [77]. While such obstruction-focused strategies can mitigate harm, they miss opportunities to intervene during the formative stages of toxicity, leading to two key limitations: (1) delayed actions cannot undo the experiential damage already inflicted, and (2) mute may shield players from harmful content, but risks disrupting the tactical coordination essential to MOBA gameplay.

Recent studies have advocated for preventive interventions and prosocial design[75–77], aiming to reduce the likelihood of toxic outbreaks by optimizing the game environment and fostering constructive interactions. For example, Bongaards et al. [8] proposed a matchmaking optimization system that allows players to select preferred match partners, thereby proactively reducing exposure to potentially toxic encounters. However, in-game preventive interventions remain underexplored, partly because their development requires a detailed understanding of the real-time evolution of aggressors' states—specifically, how cognitive evaluations, emotional fluctuations, and interpersonal conflicts accumulate into aggressive actions[19, 77].

To advance current research, our study examines toxic behavior from the aggressor's perspective, combining qualitative interviews with situational recall to trace its trajectory, from triggering events, through dynamic cognitive and emotional changes, to the enactment of toxic actions. This approach offers a deeper understanding of the mechanisms underlying toxicity in MOBA games and provides empirical evidence to inform contextual, early-stage interventions. Such interventions may help reduce the incidence of harmful behaviors while preserving the communication and collaboration essential for effective team play.

3 Study Context: League of Legends

League of Legends (*LoL*), developed by Riot Games and released in 2009, is a team-based multiplayer online battle arena (MOBA) game in which two teams of five compete for victory on a symmetrical map called Summoner's Rift [22]. Players must advance along lanes, gather resources, destroy defensive towers, and ultimately break the enemy Nexus (team bases) [22]. The game features hundreds of characters, each with unique abilities, and players need to boost their economy by defeating opponents and achieving objectives, which can be used to purchase items and strengthen their characters. Each match lasts approximately 25–40 minutes, with clearly defined roles (Top, Middle, Bottom, Jungle, Support), and gameplay is highly dependent on real-time communication and team coordination [47].

As of 2025, *LoL* boasts over 130 million monthly active players [17], making it one of the most influential MOBA games worldwide. However, its highly competitive and adversarial nature, especially in ranked modes, also leads to frequent toxic behaviors, such as

flaming, intentional feeding, AFK (away from keyboard), and excessive signal spamming (*e.g.,* pings) [40]. These behaviors undermine team coordination, escalate conflicts, degrade player experience, and contribute to player attrition [20, 21, 32]. Although Riot Games has implemented many toxicity mitigation features, like reporting systems, automated detection algorithms, and penalty mechanisms [2, 40], toxicity remains widespread, making *LoL* a common and ideal context to study toxic behaviors in MOBA games. Numerous studies have explored toxic behaviors in *LoL*, focusing primarily on their types [40], causes [24], effects [49], toxic chat detection and mitigation strategies [13]. In line with previous work, we examine *League of Legends* to advance understanding of the mechanisms and trajectories of toxic behavior, offering empirical insights to inform HCI research and the design of toxicity mitigation in MOBA games.

4 Methodology

We adopted a qualitative interview approach to explore, from the aggressor's perspective, players' perceptions of toxic behavior and transformation and evolution of the inner state of toxic behaviors throughout the gameplay. To ensure that the interview design aligned with the research objectives, we conducted pilot studies with four experienced *League of Legends* players to iteratively refine the interview structure, materials, and procedural details. The final interview protocol consisted of three steps: material collection & preparation, retrospective think-aloud (RTA), and emotional journey visualization (EJV).

Figure 1 illustrates the overall structure of the interview procedure. The combination of RTA and EJV ultimately produced a time-based, multi-perspective representation of participants' toxic experiences. Figures 2 to 7 presents an overview of a journey map. The remainder of this section presents the detailed implementation of the three key interview steps, which were informed by relevant literature [28, 72] and validated through four rounds of pilot studies. Following that, we describe the participant recruitment and elaborate on the data analysis method. Finally, we discuss several potential ethical considerations for our study.

4.1 Interview Process

This study received institutional IRB approval and obtained the consent of each participant. All interview sessions were conducted online via video conferencing software *Tencent Meeting* [1], with the virtual collaboration platform, Miro [2], used to organize and present interview materials and document outputs.

4.1.1 Material Collection & Preparation (Figure 1A). Before the actual study, participants were informed of the study's background and the requirements for interview-related materials. Upon confirming their consent to share personal match data, these players were instructed to use screen recording software during regular gameplay sessions to capture the entire game process (including the Ban & Pick [3] phase) as well as all in-game text-based chat interactions, and upload the footage with self-perceived acting as a toxic behavior initiator to us immediately afterward. All participants were prompted to play as naturally as possible, ignoring the recording when playing the game. They needed to confirm that the gameplay reflected their typical gaming behavior and stated in the submission form, to ensure ecological validity. We compiled established classifications of toxic behaviors in the existing literature [40, 43, 45] to generate the classical toxic behavior (see Appendix A). Three researchers then verified if the match recording met the inclusion criteria (i.e., completeness of data and the participants acted at least once as initiators of a classic toxic behavior). During participant screening, our goal was to verify eligibility rather than fully annotate videos. Three researchers independently went through and flagged representative potential toxic behaviors (e.g., Insulting and verbal abuse). Then, researchers discussed flagged cases in a regular meeting to decide whether these cases met the requirements, and only when researchers all agreed could the participants be invited to the next interview. When encountering ambiguous cases where consensus on the specific toxic category could not be reached (which is expected given individual differences in how toxicity is perceived [6, 32]), we used majority voting to resolve conflict. In the very rare case of three-way disagreement, the first author, who is more familiar with *League of Legends* with more than eight years of experience, made the final call on the labels later based on participants' subjective accounts in the interviews. Once verified, we scheduled their interviews within three days after the corresponding gameplay to minimize memory distortion. The final set of valid recordings had a total duration of **9.15** hours.

Before each formal interview, three researchers thoroughly reviewed the associated game recording and annotated all instances of potentially toxic events in which the participant was directly involved as an aggressor. The three researchers independently coded the acted and received toxic behavior based on the Appendix A. In addition, they also annotate major game events relevant to the participants, such as Baron/dragon loss (a neutral resource), participation in team battles, and deaths by opponents, thereby reducing the likelihood of overlooking key points during the interview. During the regular meeting, each researcher in turn presented their own annotations while the other two commented. This process allowed us to complement and filter individual observations, reducing omissions and overly subjective judgments. We discussed annotation conflicts, mainly around classifying initiated and received toxic behaviors, given that personal experience can lead to divergent interpretations of the same behavior. Considering our goal in the pre-annotation stage was to avoid missing important toxic incidents, we again used majority decisions or, in three-way disagreement, the first author's temporary judgment, while placing particular weight on participants' own interpretations during the subsequent interviews. We marked these toxic touchpoints and key game events both on the video timeline to assist in the subsequent RTA process and on a gameplay timeline plotted in Miro to facilitate the later EJV activity.

4.1.2 Retrospective Think-Aloud (Figure 1B). Our objective in this step was to reconstruct participants' cognition and emotional states when exhibiting or experiencing harmful behaviors during gameplay, as well as to identify their underlying motivations, toxicity evaluation criteria, and contextual factors associated with these

[1] https://meeting.tencent.com/
[2] https://miro.com/
[3] The Ban & Pick phase refers to the pre-game stage in competitive multiplayer games where teams alternately ban certain champions from use and select those they will play, strategically shaping the upcoming match.

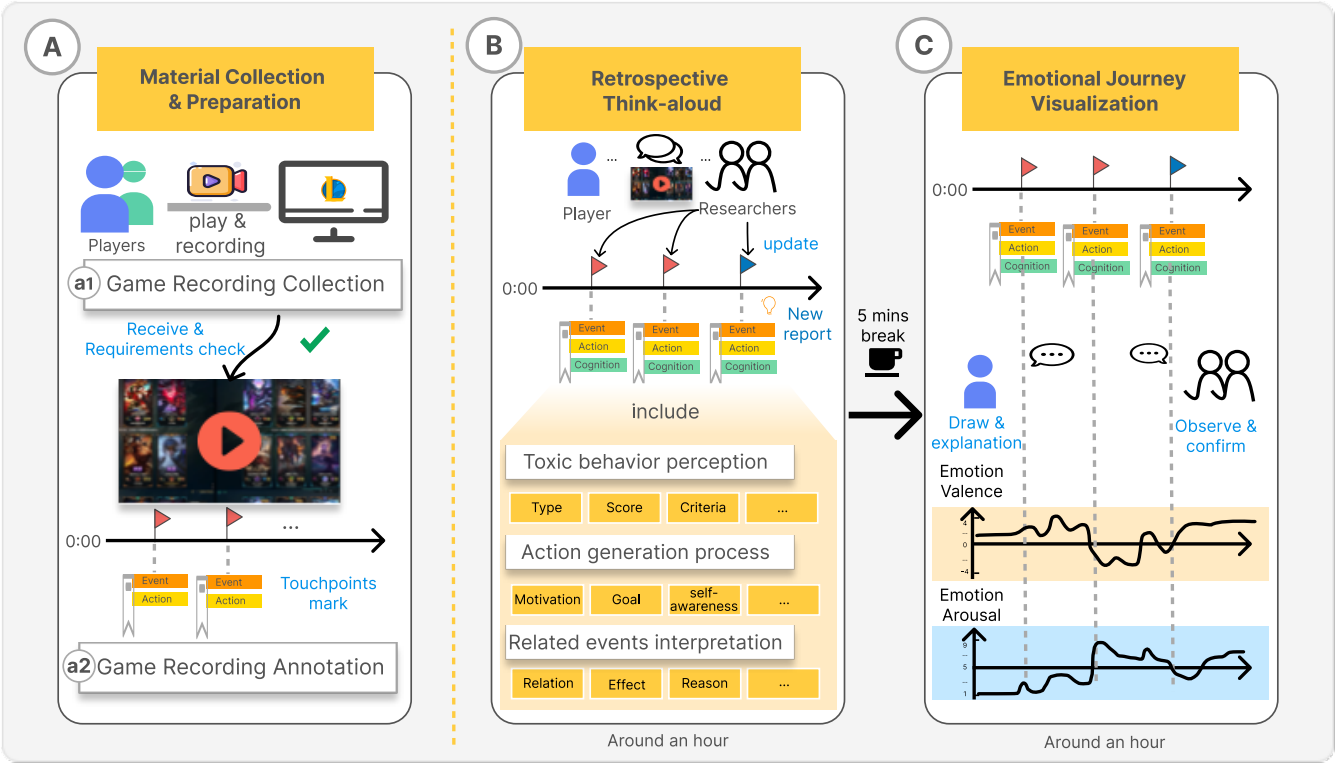


Figure 1: Interview process. The figure depicts our interview procedure. In Material Collection & Preparation (A), players self-record gameplay videos (a1) and submit them to researchers, who validate eligibility and pre-annotate key *touchpoints* (a2). In Retrospective Think-Aloud (B), researchers and players engage in in-depth exploration of both annotated and newly reported events, iteratively updating records (around 1h). After a brief 5 min break, in Emotional Journey Visualization (C), players plot *emotional fluctuation curve* based on the temporal trajectory of recorded events and self-disclosures. Throughout this stage, both parties collaboratively interpret and confirm curve patterns (around 1 hour).

behaviors. We adopted the RTA approach for two primary reasons. First, toxic behaviors tend to emerge unpredictably during gameplay, making it difficult to conduct on-the-fly interviews. Second, prompting players to express their thoughts during the match would increase cognitive load and potentially disrupt the normal flow of play [59]. In contrast, the retrospective think-aloud (RTA) method allowed matches to proceed uninterrupted while still enabling in-depth post hoc reflection.

Operationally, after obtaining informed consent, participants provided a brief overview of their general impressions and emotional responses to the recorded match to establish contextual grounding, which helps serve as a cross-validation statement, reducing the emotional impact received during the recalling process. When reviewing the Ban & Pick phase of the game, they were asked to recall and elaborate on their match objectives, initial states of mind (*e.g.*, emotion, focus, expectations, etc.), and teammate interactions at the beginning of the game, establishing context for the main game session. Throughout video playback, researchers intermittently confirmed participants' subjective experiences as unfolded in the footage. Whenever participants indicated that an event triggered a significant cognitive or emotional change, or when pre-annotated key incidents (particularly those involving potential toxicity) appeared, the segment was replayed as necessary and probed for further detail. Participants were asked to describe their immediate impressions and emotions, followed by a description along multiple dimensions, including motivation, goals, self-awareness, emotional trajectory, and causal reasoning. For toxic behavior episodes, participants were asked to rate the perceived severity of the toxic behavior towards themselves and others separately using a seven-point Likert scale (1 = low toxicity; 7 = high toxicity), provide a rationale for their rating, and discuss causal links to previous events as well as potential counterfactual scenarios.

Throughout this process, two other researchers, acting as assistants, maintained a real-time update of the event timeline on the Miro platform, ensuring that newly reported touchpoints and details were chronologically integrated into the Journey Map. Upon completion of the video review, participants verified and refined the annotated timeline. This procedure produced the scaffold of a player journey map that captures key gameplay events, toxic behavior touchpoints, and the associated players' toxicity perception information, which was used to support subsequent player emotion visualization and behavioral analysis. To mitigate interview fatigue,

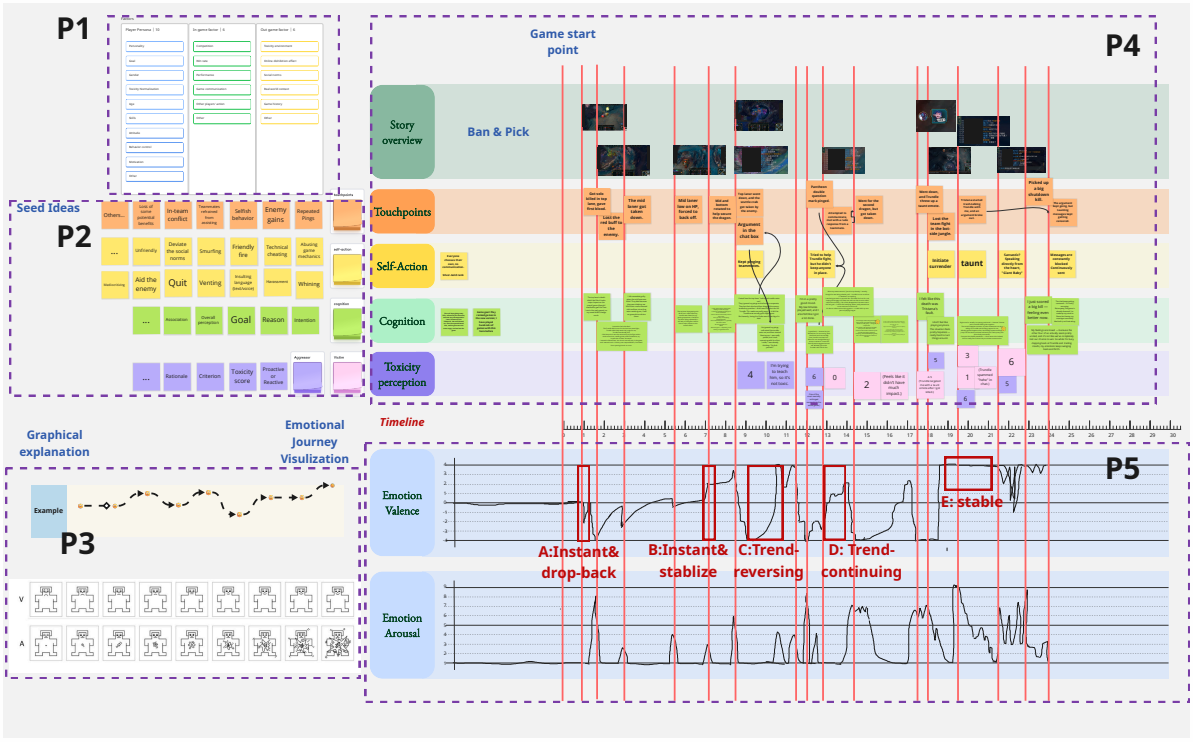


Figure 2: Journey Map Example. The figure presents an example of our *Journey Map* construction, which includes five parts: P1: potential factors, P2: seed ideas, P3: Graphical explanation, P4: Storyboard, and P5: Emotional Journey Visualization. P1, P2, and P3 are provided with prepared prompts and example perspectives to guide participants toward deeper, multidimensional elaboration. In P4, we document key situational details or player self-reflections across five parallel tracks—*Story Overview*, *Touchpoints*, *Self-action*, *Cognition*, and *Toxicity*—arranged chronologically from the initial *Ban & Pick* phase. In P5, participants then plot two temporal curves beneath the event timeline, capturing fluctuations in *Emotion Valence* and *Emotion Arousal*. Highlighted part in P5 are common fluctuation patterns observed across sessions.

participants took a five-minute break after the RTA session before proceeding to the next step.

4.1.3 Player Emotional Journey Visualization (Figure 1 C). The purpose was to characterize the temporal dynamics of participants' emotional valence (−4 to 4, where −4 = extremely negative, 4 = extremely positive) and arousal (1 to 9, where 1 = extremely low arousal, 9 = extremely high arousal) levels, by sketching on the player journey map. The valence and arousal scales were derived from the dimensional emotion model [9, 16] to jointly represent emotional polarity and activation intensity. Instead of recording discrete values only at key event nodes, participants were asked to draw continuous lines across the match timeline to trace state dynamics and highlight the changes, if any, at key touchpoints and associated events during the gameplay. This allowed us to capture and analyze user experience over time rather than in isolated moments, providing valuable context for understanding the evolving emotional, cognitive, and interactional states that shaped players' game experience and the emergence of toxic behaviors.

After participants returned from the 5-minute break, researchers first explained the conceptual meaning and scaling method of each indicator. Players then revisited key events in chronological order along the journey map timeline and plotted wave-like curves for emotion valence and emotion arousal. In this process, they highlighted curve slopes and fluctuations that indicated their psychological states at each event, verbally explaining what they encountered and how they arrived at the current level. Researchers interpreted the curves in real time, repeatedly confirming with participants that the drawings authentically reflected their subjective experiences, especially when a notable or atypical change occurred, and making immediate corrections if necessary. To reduce biases introduced by retrospective self-report (*e.g.*, just normal venting) and keep participants' accounts aligned with what actually occurred, the researchers also performed a second round of cross-validation for the same incident during the visualization process, comparing participants' earlier notes with their current drawings and descriptions. Whenever inconsistencies were identified, participants were asked to further clarify and verify the accuracy of their accounts. This dual-description design, where multiple representations of the same incident serve as mutual references, helps mitigate data biases caused by overly emotional or post-hoc reinterpretations when reviewing the gameplay video. The final sketches, combined

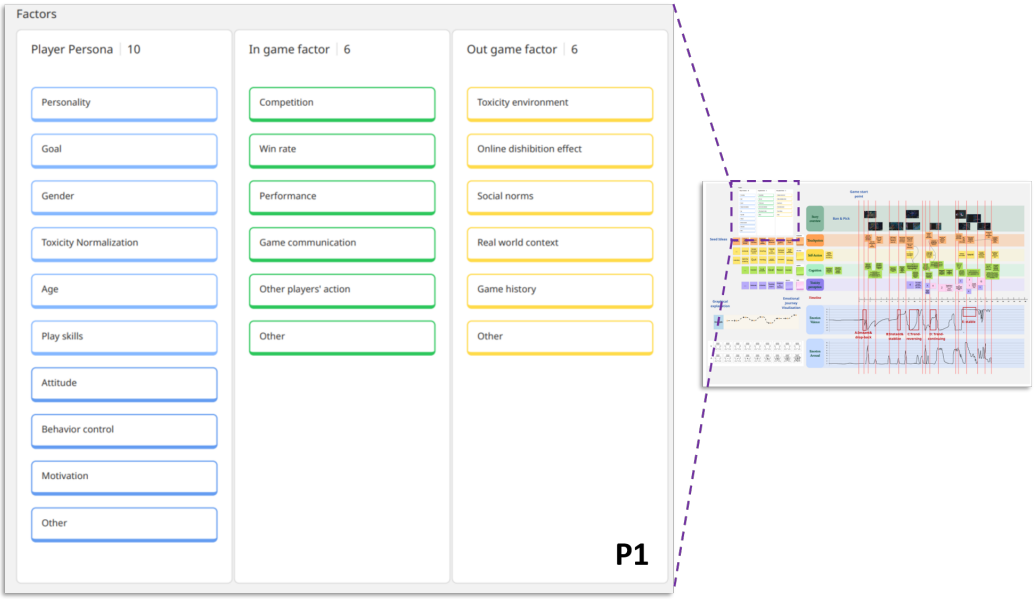


Figure 3: Enlarged version for potential related factors. (P1)

Figure 4: Enlarged version for seed ideas. (P2)

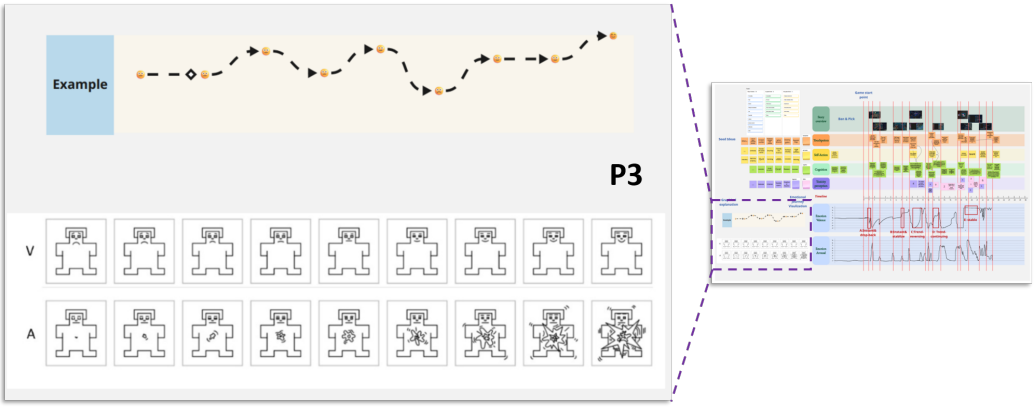


Figure 5: Enlarged version for graphical explanation. (P3)

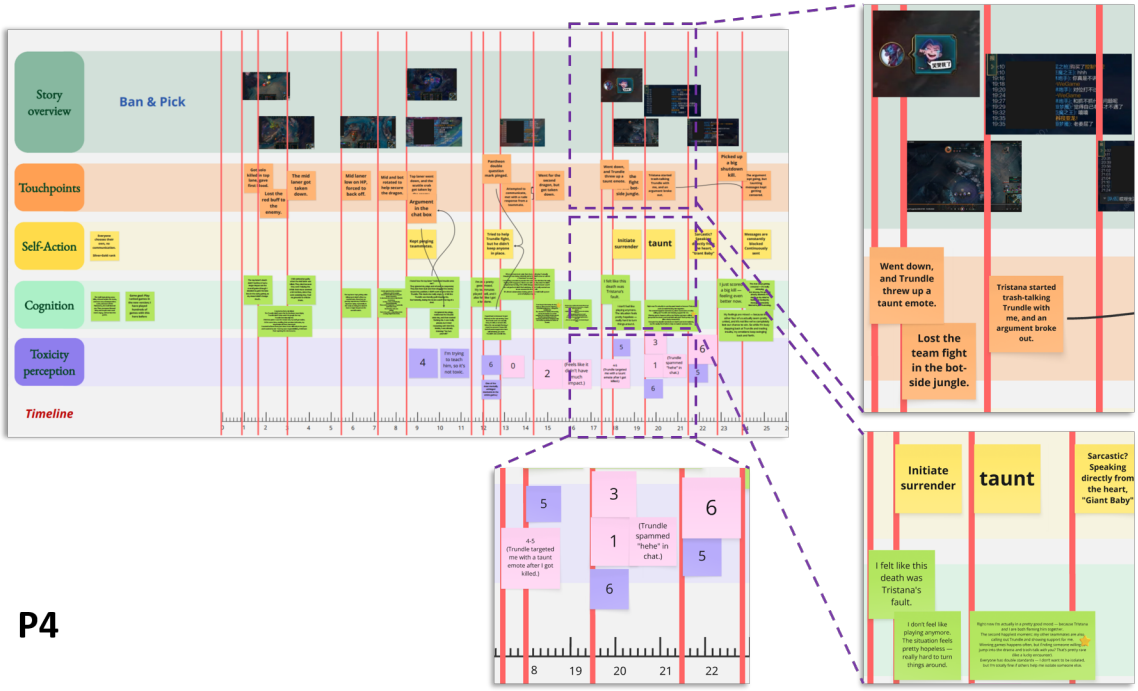


Figure 6: Enlarged version for annotation storyboard. (P4)

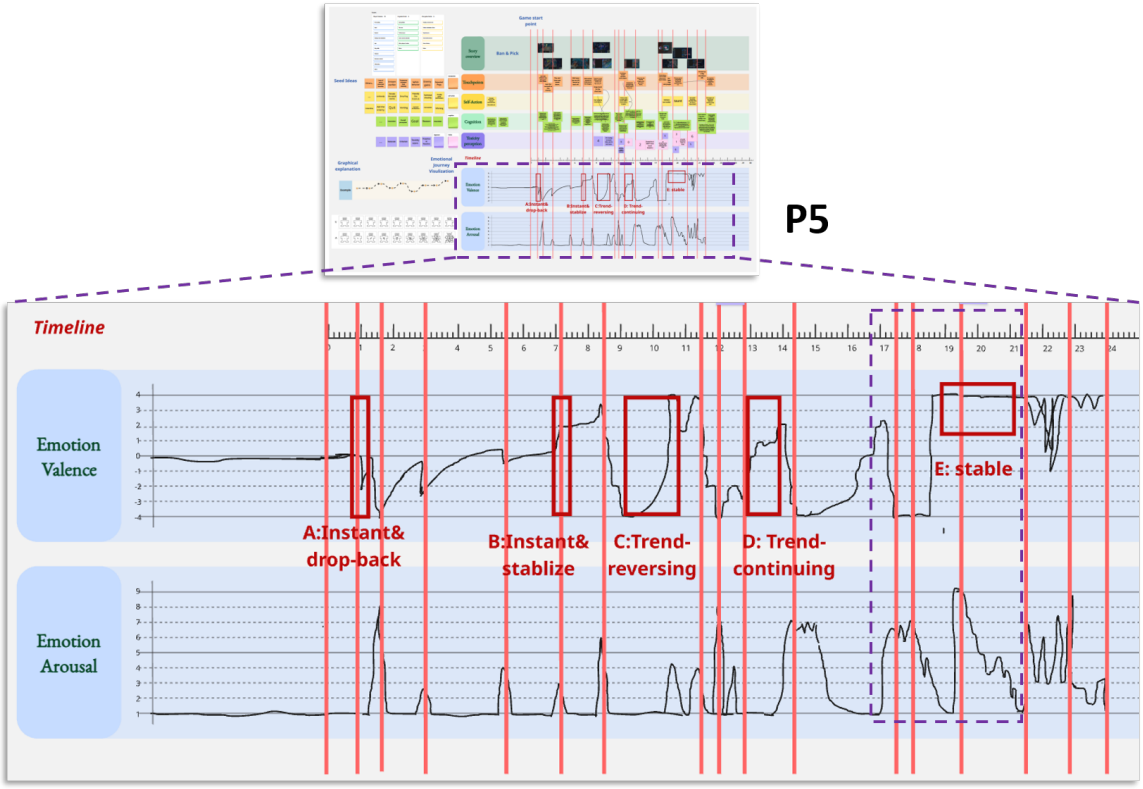


Figure 7: Enlarged version for emotional journey visualization. (P5)

with participants' verbal accounts, yielded essential sequential data to dissect the development of toxicity in gameplay.

Each formal interview lasted approximately two hours, consisting of around one hour for **Retrospective Think-aloud (RTA)** and one hour for **Emotion Journey Visualization (EJV)**, with a five-minute break between to mitigate potential cognitive fatigue. Upon completion of the entire study, each participant received 120 RMB (equivalent to 17 USD) as compensation for their contribution.

4.2 Participants

Participants were recruited through advertisements posted on multiple social media platforms, which clearly stated the inclusion criteria: (1) participants must be active players of the *League of Legends* (*LoL*) PC version; (2) they must regularly participate in the ranked game mode and currently hold an active ranked position in the game to ensure that the data reflect their normal gaming state; and (3) they agree to record their gameplay during regular game sessions in their own environments and submit the footage. After reviewing our provided definition and common examples of toxic behavior, participants submitted corresponding gameplay videos when they self-observed similar behaviors. Three researchers reviewed the player-submitted gameplay recordings to verify that each participant had initiated at least one classical toxic behavior (Appendix A) as an aggressor. In total, **18** eligible participants enrolled in our study, with a self-reported gender ratio of 13:5 (comparable to the approximate 4:1 gender ratio among the broader *LoL* player base [39, 47]). The participants' age ranged from 20 to 31 years (mean=25.0 years, SD=3.0 years) and their *LOL* experience ranged from 2 to 15 years (mean=8.7 years, SD=3.5 years). Participants represented a range of Solo queue ranks, including 1 Iron, 5 Silver, 5 Gold, 4 Platinum, 2 Emerald, and 1 Master, and brought diverse lane cases (1 Top, 4 Jungle, 3 Bottom, 10 Support). Overall, our sample encompassed a wide spectrum of players with varying experiences and skill levels.

4.3 Data Analysis

We recorded all interview sessions with consent through video, resulting in approximately 39.5 hours of video data. All recordings were first transcribed automatically through Tencent Meeting built-in transcription and then manually checked by the researchers to ensure accuracy. We analyzed the data using Inductive Thematic Analysis [26, 60], which offers flexibility in uncovering the nuances of the data and enables the identification of subtle variations within the data [10, 26, 60]. We followed recommendations for rigorous inductive thematic process, carefully documented procedures, peer debriefing, and triangulation, which emphasize transparency, team-based reflexivity, and systematic cross-checking of interpretations [11, 60], providing an appropriate form of analytic rigor without necessarily relying on a single intercoder reliability coefficient, and therefore we did not compute formal IRR statistics (e.g., Fleiss's kappa). We chose to provide a more fine-grained account of our analytic procedures in order to make the logic of our interpretations and the grounding of our themes in the data as transparent as possible [10, 11, 54]. At the same time, given the relatively small number of toxic incidents observed in our sample, we did not conduct any quantitative analyses (*e.g.*, correlations between codes or tests of statistical significance for different types of impact).

Following the standard thematic analysis process [3, 10], three researchers participated in the analysis, including the main interviewer and two assistants who attended all interviews.

First, they familiarized themselves with the data by repeatedly reviewing the recordings and then independently coding the transcripts. **Next**, in the initial phase, each researcher focused on segments that addressed the research questions, such as players' self-defined criteria for judging toxicity, their thoughts about the game context before acting, and their reported psychological journey. These codes were derived from both the interview transcripts and the annotations on the player journey maps. In parallel, we systematically used the Miro-based emotion visualizations. For each critical event in the transcript, we located the corresponding point on the journey map, examined the valence and arousal curves (including initial and end points and the slope of change), and used these patterns to interpret the intensity and development of emotional shifts around toxic behaviors. As part of triangulation, coders explicitly checked whether transcript content, video recordings, and journey map annotations converged for key moments; when they did not, the team revisited the materials together and adjusted or memoed the coding accordingly. Beyond RQ-related segments, each researcher also marked potentially valuable but unanticipated observations (e.g., players' thoughts during the ban–pick phase or interpretations of general game events without/resisting toxic behavior). Fig. 6 exemplifies how to read the journey map: after losing a teamfight, the participant is taunted by a teammate, initiates a surrender vote, and then joins the taunting exchange. In the interview, the participant reported perceiving the match as unlikely to be recoverable and interpreting the teammate's behavior as taunting, which shifted his stance toward a more offensive orientation and an immediate intention to disrupt the teammate's emotional state by provoking. The participant also described disengaging from winning, feeling the taunting interaction more interesting than competitive play. The emotion curves in Fig. 7 corroborate this shift: valence drops sharply and arousal spikes after the lost fight, then both rise rapidly to a peak as attention moves from gameplay to the taunting exchange. **After completing substantial portions of initial coding**, the team then met regularly to compare codes, resolve disagreements, and iteratively refine the codebook—merging overlapping codes, clarifying definitions, and adding new codes as needed. We sought to reduce subjective bias by grounding our interpretations as closely as possible in observable, contextualized behavior. For example, when interpreting ambiguity in ping semantics and sarcasm, we distinguished cooperative versus hostile uses of similar signals by examining how players themselves framed the incident (e.g., as a reminder or request for help versus as a complaint or challenge) and by cross-referencing these accounts with in-game actions and chat logs. **After that**, we grouped these refined codes into broader categories and themes, repeatedly checking them against the multimodal data to ensure that the resulting themes accurately reflected participants' experiences. **Finally**, we organized the derived themes and reported them using representative examples, which together constituted our findings. This stepwise, collaborative, and iterative procedure ensured the reliability and depth of the thematic analysis.

Participant ID	Gender	Age	LoL Exp. (years)	Role	Rank	Match Result
1	F	29	11	Support	Platinum	Lose
2	M	28	10	Top	Silver	Win
3	M	25	10	Support	Gold	Lose
4	M	21	3	Jungle	Gold	Lose
5	M	26	11	Support	Iron	Lose
6	M	23	11	Jungle	Platinum	Lose
7	M	28	15	Jungle	Gold	Lose
8	M	23	6	Support	Emerald	Lose
9	M	31	12	Support	Platinum	Win
10	M	20	10	Bottom	Platinum	Lose
11	M	23	9	Jungle	Silver	Lose
12	M	22	4	Bottom	Gold	Win
13	M	23	11	Support	Gold	Lose
14	M	28	10	Support	Silver	Lose
15	F	25	5	Support	Master	Lose
16	F	27	9	Support	Emerald	Lose
17	F	24	2	Support	Silver	Lose
18	F	24	7	Bottom	Silver	Lose

Table 1: Demographics of participants. "LoL Exp." refers to the number of years participants have played League of Legends.

To illustrate this process, consider one case in which a player engaged in hostile chat toward teammates after a failed team fight. From the video, we extracted a series of negatively valenced, insulting chat messages, while the transcript showed the player describing like "obviously poor performance statistics," "very low quality play," and "rather harsh language," alongside comments that the game state seemed very unfavorable. On the journey map, the corresponding timestamp displayed minimal valence and a sharp and high arousal increase, aligning with the player's report of heightened frustration. Subsequently, because there was no player interaction yet, and the negative affect before appeared to persist into the next episode, we coded that episode as a subsequent single-instance toxic behavior, after cross-checking consistency in the player's evaluations of the game situation and teammates before and after this toxic action. Finally, these initial codes were consolidated into higher-level themes.

4.4 Ethical Consideration

This study involved retrospective reviews of gameplay from the aggressor's perspective, which posed two main risks: potential retraumatization [18] and inadvertent disclosure of identifying information (*e.g.,* game IDs). Participants were fully informed of the study's purpose, procedures, and risks before recruitment. To minimize behavioral distortion, we stressed that they were under no obligation to display or suppress any behaviors, including toxicity. To preserve the authenticity of participants' recall in the interview, we did not edit the gameplay recordings to obscure toxic content or player IDs, which are already visible in the game. Instead, to safeguard participants' mental well-being, we carefully managed our tone and wording. For example, we referred to champion rather than player IDs to reduce personalization, and we avoided repeating toxic utterances, using indirect references (*e.g.,* "this statement") and cursor pointing to indicate specific content. Participants are allowed to skip questions or withdraw at any time, and the interviewer monitors their well-being throughout, adjusting pace or omitting sensitive content if needed. While gameplay recordings may capture interactions with non-consenting players, the analysis focused on events involving the focal participants, with other players' identifiers protected. All video, transcripts, and interview data were anonymized and de-identified in the analysis report and stored data (*e.g.,* to describe role/champion rather than player IDs or blurring out the player ID in the screenshot). The study was approved by the IRB and conducted within a framework designed to keep risks controllable.

5 Results

In this section, we report on the three primary research questions as well as additional findings that emerged during the interviews. We first present aggressors' perception of their potentially toxic behaviors when reviewing their gameplay retrospectively (RQ1). Next, we examine recurring patterns of cognitive and emotional states experienced by aggressors at the moment of enacting intentional toxic behaviors (RQ2). Furthermore, we explore the development patterns of toxic behavior during the game match (RQ3). In addition, we present participants' tendencies toward duty disengagement when acting as aggressors and as victims. Finally, we report how players' preconceptions influence their behavioral decisions, and players' perception of in-game content moderation mechanism.

5.1 Perception of Potential Toxic Behaviors as Aggressors (RQ1)

5.1.1 Toxic behaviors evaluation criterion. During the Retrospective Think-aloud, participants reflected on behaviors pre-labeled by the research team as potentially "toxic" according to established

Theme	Category	Definition	Example	Mentioned times
Aggressive Features	Word Offense	Speech contains direct or implicit insults or offensive expressions toward others, based on the aggressiveness of the vocabulary used	"A bunch of id*ts (P4)"	39
	Aggressive Intention	Language or action reflects targeted, derogatory, or hostile attitudes, even when the vocabulary itself is not inherently offensive	"Always slower than enemies (P6)"	27
Negative Emotion Transmission	Negative Emotion Containment	Speech or behavior contains and conveys negative emotions such as anger, frustration, or dissatisfaction	"stop M** F** kill-stealing! (P10)"	28
	Negative Emotion Amplification	Speech or behavior provokes or intensifies others' negative emotions	"As if you know anything. (P14)"	34
Timing and Frequency	——	Timing and repetition of speech or behavior contribute to cumulative negative effects	Repeatedly ping teammates when they are killed by enemy (P13)	25
Scope and Game Consequences	——	Negative impact of speech or behavior affects individuals, the team, or directly influences the match situation	Away from keyboard (AFK) (P2)	23
Rationality	——	Speech or behavior is based on objective facts, includes explanatory elements, or occurs in a relatively reasonable context	Ping when noticing the teammate stopped operation in front of an enemy (P9)	25

Table 2: Toxic behavior severity judgments dimensions synthesized from the interview transcriptions with aggressors. Aggressors will consider a comprehensive assessment (rather than relying on just one theme) when rating the severity of their toxic behavior.

taxonomies [40, 43, 45, 52], as well as behaviors they personally identified as toxic during gameplay. Thematic analysis revealed that aggressors evaluated the toxicity of their own actions across five interrelated dimensions: *aggressive features*, *negative emotional transmission*, *timing and frequency*, *scope and game consequences*, and *rationality*. This multidimensional standard extends the three-factor framework proposed by Laato et al. [45] for *StarCraft II* [4] — directly observed, in-game contextual, and extraneous factors — by refining it into five specific dimensions (seven including sub-dimensions). It reveals that toxicity evaluations in gameplay are both subjective and context-dependent, demanding an integrated perspective. In the following, we elaborate on each criterion from the aggressor's perspective (Table 2).

► **Aggressive Features:** Aggressive Features refer to speech or behavior expressing hostility or offensiveness toward others. Based on the content and intention, they are divided into *Word Offense* and *Aggressive Intention*. *Word Offense* captures the extent to which players' speech contains direct or implicit insults and offensive expressions, often "harsh language" with community-specific slurs [27] targeting intelligence, ability, personality, or family. These overt expressions have clear semantics, require little contextual inference, and tend to produce immediate negative impact, typically rated as at least moderately toxic in our participants. In contrast, *Aggressive Intention* assesses how strongly language or actions convey targeted, derogatory, or hostile attitudes even without explicit vulgarity. Here, context, delivery, and perceived intent are decisive: seemingly neutral remarks like "Always slower than enemies" can be moderately toxic when reproachful (P6). Such covert but high-intent expressions, described as having "covert lethality" (P7), can evade automated moderation while still undermining morale and team cohesion [6].

► **Negative Emotional Transmission:** Negative Emotional Transmission refers to speech or behavior that conveys or spreads negative emotions such as anger, frustration, or dissatisfaction, and is divided into Negative Emotion Containment and Negative Emotion Amplification. Negative Emotion Containment assesses how strongly speech or actions directly express negative emotions—often through strong emotional particles (*e.g.,* tmd), rapid

[4] https://starcraft2.blizzard.com/en-us/

delivery (e.g., repeated avatar clicks, "?" pings), and usually co-occurrence with harsh language or explicit aggressive intent, which together heighten perceived toxicity (e.g., "Stop M* F* kill-stealing!" (P10)). Purely emotional outbursts without personal attacks were typically rated as less toxic, as not seen as targeted harm (P10). Negative Emotion Amplification evaluates how much aggressors perceive their speech or actions as provoking or intensifying others' negative emotions, often via sarcasm or exaggerated criticism of failures, which can significantly escalate toxicity by aggravating interpersonal conflict even without vulgarity (*e.g.,* "Still thinking you're an unrecognized genius? Aww, feeling all hurt?" (P7)).

► **Timing and Frequency:** This dimension highlights that the timing and repetition of speech or actions shape perceived toxicity. Aggressors often noted that making sharp or aggressive remarks when others are emotionally unstable, have just made mistakes, or when the team is disadvantaged is more likely to be seen as malicious, leading them to judge their own behavior as more destructive. Sensitivity to timing thus became a key reference in self-assessment—for example, one participant (P11) described sending a thumbs-up emoji or pinging a teammate right after they died as "taking advantage of the moment when they're most annoyed." Frequency was also crucial: prolonged complaints, revisiting past mistakes, or persistent pinging were said to create a sense of "oppression," like "nagging" or "a mosquito buzzing around your ear" (P6). In contrast, infrequent minor behaviors (e.g., a single complaint without harsh language or strong emotion) were usually rated as low toxicity, as aggressors felt such acts would not attract lasting attention (P11).

► **Scope and Game Consequences:** This dimension reflects the degree to which the reach and in-game consequences of speech or actions contribute to perceived toxicity. Participants noted that explicitly targeting an individual heightened the sense of personal attack, whereas addressing the entire team often reduced personal identification and lowered toxicity ratings (P1, P5). However, when such conduct was perceived to damage collective morale, aggressors reported raising their own toxicity assessment (P7). Direct in-game impact also served as a key criterion: behaviors such as going Away From Keyboard (AFK) were described as "directly causing imbalance in the game" (P6), harming teammates' experience, and diminishing the chance of victory, thus regarded as highly toxic. Conversely, when an action was seen as having no tangible in-game effect, such as not disrupting operation, attackers tended to downplay its toxicity or frame it as "just normal venting" (P11, P13).

► **Level of Rationality:** This dimension examines how much speech or actions are grounded in facts or reasonable justification, shaping aggressors' perceived toxicity. Participants often sought "legitimate" reasons to downplay their toxicity, insisting they were reacting to abnormal behavior rather than attacking "for no reason". They tended to rate their behavior as less toxic when criticism was supported by observable facts (e.g., repeated mistakes, rule violations, intentional misconduct) and delivered without abusive language, framing it as a "reasonable emotional response". For example, P1 justified her complaints by saying, "My teammates never remind me that the enemy is coming, so I kept getting killed again and again." However, they also recognized that justification does not guarantee mildness: when rational points were accompanied by exaggeration, sarcasm, or public humiliation, they felt toxicity increased, describing this as "having a reasonable point but going too far," and still judging such behavior as highly toxic for exceeding social norms.

5.2 Internal State Patterns at Toxic Behavior Emergence (RQ2)

Based on general aggression model [4], we further analyzed players' internal state when engaging in toxic behaviors from the perspectives of cognition and emotion. A thematic analysis of players' self-narratives during intentional toxic behaviors indicates that toxic behavior is jointly driven by a hierarchical cognitive structure and escalated emotional processes. Figure 8 shows the detailed hierarchical cognitive structure.

5.2.1 Hierarchical cognitive structure. Based on our data analysis, we propose an interpretive three-layer structure observed in participants' accounts, which consists of the **Perception Layer**, the **Orientation Layer**, and the **Intention Layer**. In the following, we detail the components within each of these layers and explain the influence among them.

► **Perception Layer:** The perception layer refers to the aggressor's immediate and intuitive understanding of the current game context. Depending on the object of perception, it can be divided into three categories: situation perception, event perception, and player perception.

Situation perception reflects the aggressor's sense of overall game progression (e.g., game phase, team strength, expected trajectory), providing a macro context for interpreting events. Event perception represents the aggressor's analysis of particular game events or interactions, typically occurring during post-event attribution of responsibility and assessments of rationality. Player perception refers to the aggressor's overall impression of others, such as their gaming skills or personality traits, which evolve dynamically as new events are observed throughout the game.

These three types of perception are interrelated and mutually influential. Situation perception immediately shapes how events are interpreted, while accumulated event perceptions continually update the aggressor's sense of the situation. For example, a teammate's death may be seen as a "minor issue" when ahead (P6) but "unacceptable" when behind (P8), and repeated failures in resource contests can lead to believing "victory is out of reach" (P1). Situation and event perceptions jointly inform player perception: when evaluating events, aggressors attribute responsibility and judge rationality, intensifying negative impressions when blame is placed on others and easing them when a reasonable explanation exists. For instance, a missed warning that leads to death can cause distrust, but if seen as a first-time oversight, the negativity may be mitigated (P1), indirectly moderated by situation perception. Player perception feeds back into situation and event perceptions: losing confidence in a teammate's skills can make the game feel "unwinnable" (P4), and the same ping may be interpreted as "normal complaining" before a negative impression forms, but later as "intentional provocation or taunting" once hostility is presumed (P11).

► **Orientation Layer:** The Orientation Layer refers to a persistent attitude moderated by perception layer, which encompasses the habitual ways of thinking that guide players' specific behaviors in

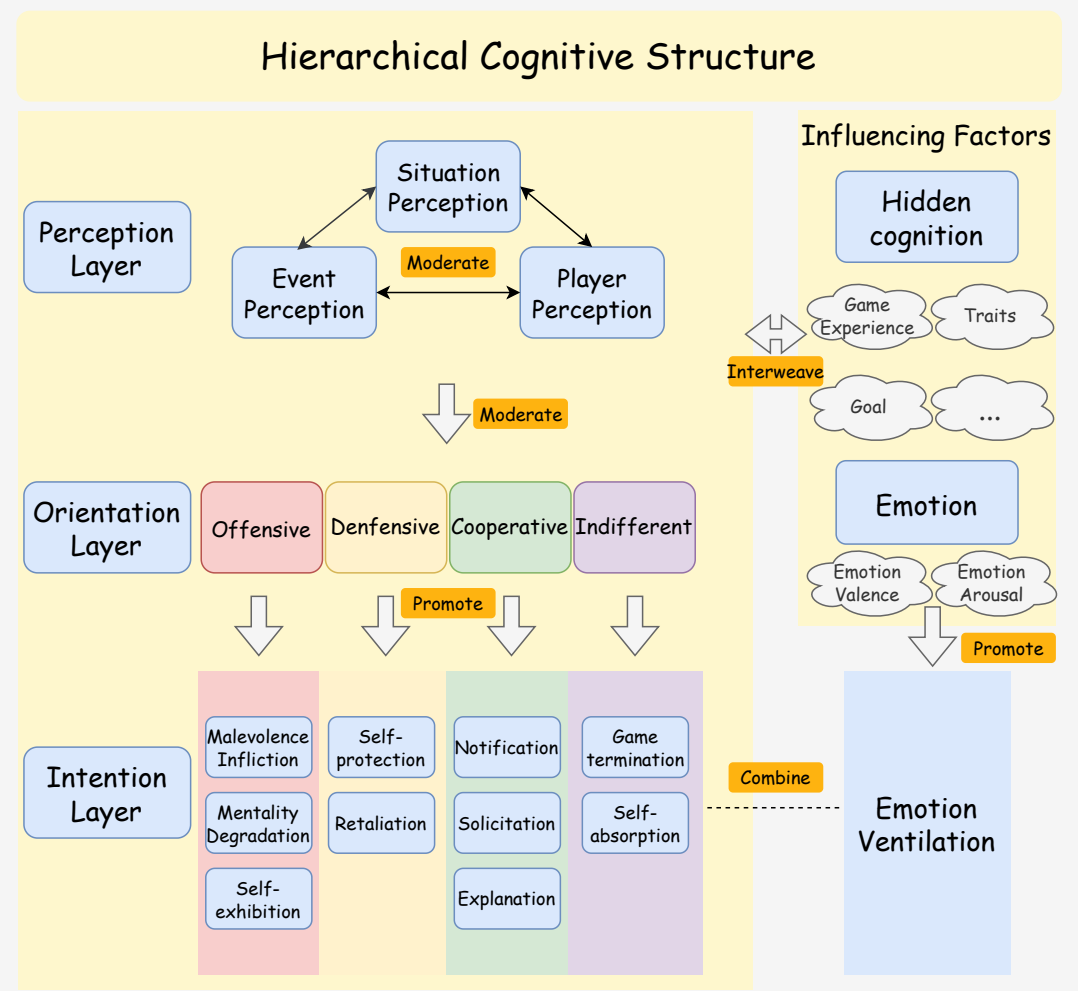


Figure 8: Hierarchical cognitive structure. The figure illustrates our three-layer cognitive framework for toxic behavior: the *Perception Layer*—comprising *Situation Perception, Event Perception,* and *Player Perception*—features reciprocal influences that collectively modulate the *Orientation Layer* (*Offensive, Defensive, Cooperative, Indifferent*). Each orientation promotes distinct toxic behavior intentions within the *Intention Layer,* alongside *emotional ventilation* facilitated by *Emotion*. Both *Hidden Cognition* and *Emotion* function as external influencing factors, shaping all three layers and ultimately informing the manifestation of toxic behaviors.

the game. The Orientation Layer mainly includes four types of tendencies: Offensive Orientation, Defensive Orientation, Cooperative Orientation, and Indifferent Orientation.

Offensive orientation reflects that the aggressor tends to initiate toxic behavior in an active and proactive manner. This orientation is closely tied to the aggressor's negative perceptions of others. When other players are judged as "lacking teamwork," "acting abnormally," or "displaying negative attitudes," aggressors easily form a "deserving punishment" mental framework, which quickly translates into action when those players are involved in negative incidents. For example, P2 criticized an unskilled teammate for "lacking self-awareness" and being "unworthy of getting kills," while P7, after repeated verbal harassment and failed explanations, deemed a teammate "impossible to reason with" and resorted to pinging upon their death.

Defensive orientation reflects a passive tendency to respond to offense, harassment, or aggression when perceived as personally targeted or disturbing. It does not involve actively seeking conflict; rather, it emphasizes a self-protection mechanism. The goal can be to block others' attacks, defend oneself from violation, or simply retaliate in a tit-for-tat manner. For instance, P13, after a surrender suggestion was met with "shut up," replied in kind, while P18, following toxic exchanges, refocused on gameplay but vowed to curse at a teammate in retaliation.

Cooperative orientation reflects an attitudinal tendency to pursue victory through team collaboration, even when expressed via behaviors with toxic characteristics. While partly rooted in the structural demands of the game, the cooperative orientation of aggressors often centers on ensuring others align with their own coordination needs. P3 and P8 repeatedly pinged "?" to prompt teammates, with

P3 to warn of an ambush, and P8 to urge the jungle to contest resources, illustrating unilateral communication with cooperative intent but antagonistic overtones.

Indifferent orientation reflects a gradual loss of interest in match outcomes and objectives, leading to psychological disengagement from gameplay. It is often triggered when aggressors perceive the match as irreversibly disadvantageous, prompting reduced effort, frequent surrender votes, or mechanical participation in the game. For example, P13 repeatedly voted to surrender "to end it sooner and start the next one," while P10 ignored the team's retreat and continued fighting alone in enemy territory. Additionally, when aggressors perceive their teammates as "unworthy of winning," this can also trigger an indifferent orientation. For instance, after multiple toxic interactions with teammates, P2 became extremely disappointed and chose to go AFK to signal "no longer willing to help the team win."

Aggressors rarely act from a single orientation; instead, orientations co-occur and are jointly moderated by the perception layer, shaping behavioral intentions in nuanced ways. Cooperative orientations may still embed offensive orientation (*e.g.,* P3 and P8 mentioned), while the co-presence of offensive and indifferent orientations can attenuate or amplify toxicity. In some cases, indifference dampens emotional investment, softening offense—after perceiving no chance of winning, P4 was "unbothered" when a teammate went AFK, believing it would "help end the game sooner." In others, detachment fuels antagonistic amusement—having abandoned the match, P7 kept pinging to "annoy" a previously harassing teammate, finding it "amusing" to elicit frustration. These patterns suggest that orientations operate as dynamically interacting dispositions rather than isolated states.

► **Intention Layer:** The intention layer refers to players' immediate, context-dependent motivations for engaging in toxic behaviors, serving as the proximal driver between the orientation layer and observable actions. Unlike the relatively stable orientation layer, intentions are transient and do not necessarily represent long-term behavioral tendencies. We identified five common types: four aligned with orientation categories—*offensive intention*, *defensive intention*, *cooperative intention*, and *indifferent intention*—and a fifth, *emotional ventilation*, denoting a desire for emotional release independent of orientation.

Offensive intention typically arises when offensive orientation dominates in gameplay. It refers to the immediate motivation to inflict offensive impact on others, primarily consisting of malevolence infliction, mentality degradation, and self-exhibition. Depending on the focus of expression, malevolence infliction denotes the aggressor's explicit desire for the victim to perceive negative or hostile intent, ranging from complaints, doubts, accusations, to personal attacks; for instance, referring to teammates as "You're all fing animals." (P4). Mentality degradation describes the aggressor's attempt to impose psychological pressure, frustration, or emotional imbalance on the victim; for example, "Who the hell brought you into this game?" (P15). Self-exhibition refers to the aggressor reinforcing their presence and sense of superiority by belittling others or boasting after performing a highly skillful action, such as a clutch kill or escaping from danger; for instance, after killing an opponent, P10 sends "?".

Defensive intention typically arises when defensive orientation is activated in gameplay. It refers to the immediate motivation to respond to perceived injustice, attack, or potential threat, primarily consisting of self-protection and retaliation. Self-protection denotes the player's attempt to deter further targeted behavior; for instance, after being mocked by a teammate, one participant executed a remarkable kill and immediately used a taunting emote to discourage further verbal attacks (P14). Retaliation describes actions aimed at restoring a perceived imbalance; for example, "He pings me first, so of course I ping him back. as a fight back" (P4). While self-protection is generally reactive, behavior can intensify when emotions accumulate, sometimes surpassing the original provocation. However, once the perceived threat or offensive interaction ceases, both the frequency and severity of defensive behaviors typically diminish. "He stopped replying with toxic messages, so continuing to attack him would make me look overly persistent" (P18).

Cooperative intention typically arises when players perceive a need for team coordination to achieve shared objectives during gameplay. It refers to the immediate motivation to facilitate collaboration, primarily consisting of notification, solicitation, and explanation. Notification denotes the player's effort to convey key situational information to the team in real time, prompting timely responses; for example, "I saw them taking the dragon, so I immediately pinged to alert my teammates." (P8). Solicitation describes explicit requests for assistance or resources, often under adverse circumstances and emphasizing the urgency of teammate intervention; for instance, "The bottom lane is unplayable, and the jungle is just farming nearby. I really hope he comes to help soon." (P14). Explanation refers to clarifying one's actions or tactical decisions to dispel misunderstandings or stabilize the team atmosphere, such as "I told my teammate I didn't take his resources on purpose." (P5). Within toxic interaction contexts, cooperative intention may be expressed through excessive signal spamming, implicit blame, or emotionally tinged coordination demands. While the game's structure inherently facilitates team play, prior negative perceptions of teammates can intertwine with cooperative needs, producing collaboration efforts that carry antagonistic undertones.

Indifferent intention typically arises when players lose investment in team victory or overall game objectives. It refers to the decision to disengage from collective goals, primarily consisting of game termination and self-absorption. Game termination denotes attempts to end the match early, often driven by frustration or predictions of inevitable defeat; for example, "My teammates are just messing around, there's no way to win this game." (P13). This detachment may also override positive match conditions, as in P2's case, where even with a lead, they chose to go AFK because "the teammates were just too annoying; even if we won, I wouldn't feel happy." Self-absorption describes a shift toward prioritizing personal enjoyment or emotional satisfaction over team strategy, such as "They didn't care about me, so I just played for my own fun." (P10).

Emotional ventilation typically arises when players experience intense emotional agitation or an accumulation of negative feelings during gameplay. It refers to the immediate impulse to release such emotions through in-game behaviors. Unlike intentions tied to a specific orientation, emotional ventilation can occur within offensive, defensive, cooperative, or indifferent contexts, serving

as a catalyst that escalates the intensity of other toxic behaviors. For example, aggressors often amplify toxic expression by inserting swear words or expletives into their phrasing. As in P10's case, who said "Stop M* F* kill-stealing" to a teammate; here, the profanity does not directly target the other person but modifies the command to convey anger. While such expressions may not directly serve strategic goals, they provide temporary emotional relief, and are particularly likely to emerge in high-pressure, competitive matches.

► **The interweaving of emotions and hidden cognition:** Our observations indicate that both an aggressor's emotional state and underlying *hidden cognition* can shape shifts within the hierarchical cognitive structure. *Emotions* may intertwine with any toxic behavior intention—*offensive*, *defensive*, *cooperative*, or *indifferent*—amplifying the intensity of expression. *Hidden cognition* denotes pre-existing attributes such as prior gaming experience, skill level, and personal toxicity norms, which can shape the *perception layer*; for instance, P18 recalled that a teammate's accusation "reminded her of many similar past incidents," heightening distress and reframing interpretation. Such predispositions may also influence orientation and ensuing intentions: P5 noted that he "generally avoided arguing" and only responded when provoked, reflecting a predominantly responsive orientation, with most toxic acts aiming to resolve misunderstandings rather than initiate conflict. The complexity of these interrelations remains underexplored, warranting further investigation.

5.2.2 Emotion state associations with toxic behavior. We analyzed changes in emotional valence and arousal before and after events using players' self-reports and journey map curves, identifying three patterns: *stable* (minimal change, indicating low event impact) (Figure 2 E), *instant change* (rapid shift post-event, subdivided into *instant&stabilize* (Figure 2 B) or *instant&drop-back*) (Figure 2 A), and *gradual shift* (slow change toward a new equilibrium, either *trend-reversing* (Figure 2 C) or *trend-continuing* (Figure 2 D)). Players' self-rated toxicity was mapped as 1–2: low, 3–5: moderate, 6–7: high; Emotion valence as 0: neutral, 1–2: slightly positive, 3–4: positive, −1 to −2: slightly negative, −3 to −4: negative; and Emotion arousal as 1–3: low, 4–6: moderate, 7–9: high.

► **Emotion Valence:** Self-reported valence–toxicity score shows that neutral or lower valence is more often linked to medium/high toxicity, reinforcing the association between negative affect and toxic expression. Medium/high toxicity also occurred under positive valence, primarily when *indifferent* and high-*offensive* orientations co-occurred, indicating punitive rather than cooperative states (*e.g.,* "I just kept pinging him for amusement" (P11)). Positive or negative valence typically remained stable at toxicity onset, as consolidated perception states (e.g., confident lead or resigned loss) made single non-critical events may insufficient to shift affect. While neutral to mildly positive/negative states were more event-sensitive, often following *trend-reversing* or *instant & stabilize* shifts toward negativity. Although toxic triggers typically deepened negative affect, partial recovery was common without further provocation, stabilizing at slightly negative levels due to persistently low win expectations.

► **Emotion Arousal:** Most toxic behaviors emerged at moderate-to-high arousal, with toxicity level generally scaling with arousal. Toxic-triggering events often elevated arousal through *trend-reversing* or *instant & stabilize* patterns, though some cases showed arousal decreases—typically when players saw the match as unrecoverable and disengaged (P4) or deliberately avoided retaliation to refocus on objectives (P16, P18). These cases suggest that attentional shifts and psychological detachment can temper sustained arousal.

5.3 Toxic Behavior Development Trajectory (RQ3)

Analysis of toxic-behavior trajectories suggested two recurring patterns in our sample based on frequency within the same context: *persistent* behavior and *single-instance* behavior, the latter further categorized as *isolated*, *subsequent*, or *repetitive*. A *context* denotes a stable situational framework within a gameplay segment, defined by interactions around a specific theme, target, or event. Context consistency drew on multiple situational cues—temporal proximity, thematic continuity, and alignment of focus/attention—rather than a single metric. Strong cue coherence, with emotion, attention, and interaction logic extending from prior states, indicated a continuous context; otherwise, behaviors were assigned to a distinct one.

5.3.1 Persistent toxic behavior. Persistent toxic behavior is defined as repeated toxic actions within a continuous gameplay context. Persistence often stemmed from ongoing interactions, sustained by a feedback loop where negative emotions and perceptions were reinforced by others' responses or persistent in-game conditions. For example, P18's argument with a support player lasted 2.5 minutes across recalling, traveling, and farming, with valence dropping from slightly negative to negative, arousal rising from moderate to high, perceptions worsening, and toxicity escalating from level 3 to 5 ("dumb dog"). When the support stopped responding, P18 ended the exchange, not wanting to "seem too aggressive." Persistent toxicity also occurred without continuous interaction. P14 repeatedly criticized the jungle for "just farming"; receiving no reply, P14 followed to steal resources and typed "Let's see how you play," deliberately disrupting the teammate's experience. In this case, persistence was driven by one-sided affective needs, with disruption itself providing gratification and emotional release. The match stage also influenced persistence. In late-game contexts, it rarely ceased naturally, as players deprioritized gameplay and focused on conflict, often messaging during deaths or halting play. In early or mid-game, frequent events made persistence more interruptible: enemy engagements, task execution, or improved situations redirected attention to gameplay, ending toxic behavior.

5.3.2 Single-instance toxic behavior. In contrast to *persistent toxic behavior*, we define *single-instance toxic behavior* as a pattern in which only one toxic act occurs within a given gameplay context. Between such instances (if any), players' attention generally shifts back to gameplay operations, and no sustained verbal or behavioral escalation takes place. Based on the temporal interval between occurrences and the degree of contextual similarity, we classify single-instance toxic behavior into three subtypes: *isolated*, *subsequent*, and *repetitive*.

► **Isolated single-instance toxic behavior:** *Isolated* single-instance toxicity is the basic pattern, typically arising when players are focused on progression and triggered by events unrelated to prior contexts. It appeared mainly in early and mid-game phases, where farming or advancement increases new context encounters.

As triggers rarely had a decisive impact, players often maintained neutral perceptions, so toxicity remained low, often limited to habitual post-error pings like "?" (P11). Situation perception shaped emotional responses: unexpected events prompted abrupt shifts (e.g., P1 was ambushed and killed, then pinged "?"), while expected events produced more gradual reactions (*e.g.,* P1 lost a team fight under disadvantageous conditions and blamed teammates).

► **Subsequent single-instance toxic behavior:** This type of *single-instance toxic behavior* occurs after a prior toxic episode, when the player's emotional state has not fully stabilized and a new, contextually different event quickly rekindles dissatisfaction. Unlike non-sequential cases, where arousal subsides to moderate or low levels, sequential cases involve recovery interrupted by a fresh negative stimulus, causing arousal to surge again before stabilization. For example, P2 experienced "consecutive teammate deaths → not assisting him → critical mistakes → kill stealing," with interim drops in arousal followed by renewed spikes. This differs from *persistent toxic behavior*, where arousal remains elevated in the same context; here, it falls in the interval but rebounds when a new situation intensifies the partially recovered state.

► **Repetitive single-instance toxic behavior:** *Repetitive* single-instance toxic behavior occurs when players encounter contexts similar to prior negative experiences—not necessarily the same, but sharing salient features such as repeated teammate mistakes. These instances become cognitively linked, forming an accumulated negative impression (e.g., "this player always makes mistakes"), which can amplify emotional reactions upon recurrence. For example, when P10's teammate first took their kill, P10's emotion shifted to slightly negative with moderate arousal, perceiving the teammate as "lacking team awareness." After the regulation, the same event happened again, triggering a sharper drop in valence and higher arousal ("Teammates do not value me"). The first occurrence of toxicity may also exhibit this pattern, as initial incidents can lower tolerance thresholds without immediate expression; negative emotions accumulate internally, surfacing only when recurrence exceeds the threshold. For instance, P1 dismissed her first sudden death as "Teammates maybe just not pay attention" but reacted more strongly to a similar second death, perceiving teammate behavior as "abnormal".

5.4 Role-Related Nuances in Attributing Toxicity

Within a single match, players may act as aggressors, victims, or bystanders. Although severity ratings of toxic behavior remain similar across roles, role-specific differences emerge in responsibility attribution and interpretive stance. Victims often label others' actions as "unreasonable" or "groundless," expressing feelings of wrongful accusation (*e.g.,* describing insults as "purely picking a fight") (P18). Aggressors more readily view others' actions as "worthy of criticism," shifting blame to the other party (*e.g.,* faulting a teammate after a failed tower dive) (P12). Bystanders usually ignore conflicts unrelated to them, especially in fast-paced or high-stakes moments. When they do intervene, their comments may be seen as taking sides, sometimes amplifying players' sense of being wronged (*e.g.,* P6's teammate saying "he (refers to P6) only took your a bit resource, why care?") or reinforcing attacks (*e.g.,* P7's teammate joining in to blame another player). Overall, victims tend to perceive others as irrational, aggressors are inclined to criticize, and bystanders largely avoid involvement—yet interventions do not always de-escalate tension.

5.5 Cognitive Predispositions Formation

In most cases, players avoid interaction during the BP (Ban/Pick) phase, largely due to the perception that early communication is more likely to provoke conflict than promote cooperation, consistent with findings by Lee et al. [47]. However, our observations suggest that early positive impressions—formed through casual conversation or proactive assistance—can increase tolerance toward teammates' later mistakes and reduce the likelihood of severe toxicity. For example, players who found teammates "interesting" or "willing to sacrifice for others" reported avoiding personal attacks even when the team was losing (P7, P13). When such early positive expectations are violated, the resulting psychological gap may cause negative responses. P5, initially impressed by a teammate's strong performance and cooperative spirit, became openly hostile to other teammates after that same player announced they were "deliberately feeding." These findings indicate that early interactions can shape cognitive predispositions that may influence both the interpretation of others' behavior and the intensity of subsequent emotional fluctuations.

5.6 Mute is Not a Cure

League of Legends employs automated language detection and filtering to block offensive vocabulary. This distinction is necessary, as interview feedback indicates that players generally regard "offensive words" as indicative of a very high level of toxicity. While such systems prevent some explicit abuse, they have a limited impact on overall toxicity. We observed that aggressors in Chinese-language contexts frequently use homophones or abbreviations to circumvent chat filters. Also, sarcastic speech are widely recognized for their ability to provoke strong emotional reactions in victims, yet these expressions are difficult to accurately detect and filter. In addition, repeated "message blocked" notifications rarely de-escalated emotions and sometimes increased frustration, with players expressing a preference for unfiltered expression for emotion release (P2). Moreover, emotional suppression sometimes shifted toxicity from language to behavior: after being muted for verbal abuse, P10 disengaged from team cooperation, a kind of toxicity invisible to the filter yet potentially more damaging to match outcomes.

6 Discussion

Our findings describe criteria that aggressors in our sample reported using when evaluating toxic behavior, illustrate changes in cognition and emotion they associated with toxic incidents, and outline developmental patterns of toxic behavior that we observed across the course of matches in our data. We discuss how these results provide deeper insight into the evolution of toxic behavior, both in brief episodes and over the course of entire matches. Potential directions for future exploratory research are also suggested. Although our findings come solely from a small-scale dataset in League of Legends, many key situational features (*e.g.,* intense team-based competition, ranked pressure, amplified blame for teammates'

mistakes) are common across other MOBA games [50](*e.g.,* Dota 2, Honor of Kings). We therefore expect that some of the recurring patterns we observe—such as how players accumulate frustration, rationalize toxic behaviors, and draw boundaries between "jokes" and genuine attacks—may transfer within the MOBA genre. However, cross-game differences in game design, penalty schemes, community norms, and even device characteristics (*e.g.,* differences between mobile and PC devices in the ease of typing or issuing ping signals) may shape how toxicity manifests. Accordingly, we position our findings as an explorative lens for MOBA games, rather than universal claims about all titles in this genre, which should be further examined in different games and environments, like interviewing broader player populations. Based on this, we further explore how our findings can inform preventive approaches for the mitigation of toxic behavior in MOBA games, and propose corresponding intervention and game design recommendations.

6.1 Implication for Theory: Understanding Toxic Behavior in MOBA Games

Prior research on toxic behavior in online multiplayer environments has examined either aggregated patterns of negative interactions (e.g., frequency of reports, general behavioral tendencies) or static correlates such as demographics, personality traits, and in-game trigger events [36, 44, 45, 51]. Such approaches often treat toxicity as a fixed outcome—identifying who is toxic and why—while neglecting its temporal trajectory, players' latent internal states, and the mutual influence between unfolding in-game events [7, 40, 49]. Although these perspectives have yielded valuable insights into prevalence and antecedents, they lack sensitivity to in-situ game dynamics and to moment-to-moment fluctuations in aggressor decision-making. Recent research has begun to attend to the dynamics of gameplay, such as role transitions and cascading effects of toxicity Kordyaka et al. [36]. For example, Kordyaka et al. [36] have proposed a "cycle of toxicity" model in which players move between perpetrator, victim, and bystander roles over the course of a match, and in which these role transitions are linked to the fluid circulation of toxicity among players. They argue that once initial elements of toxicity are introduced into a game, they can trigger further toxic responses and thereby initiate and reinforce a self-amplifying cycle of toxicity. This model operates primarily at a relatively macro level of events and player roles, points out that toxic behavior escalates, spreads, and recurs between players across different phases of a match, and thus acknowledges the dynamic nature of toxicity in gameplay. However, it still does not provide a clear account of how, at a more micro level, players' hidden internal states emerge, escalate, or dissipate in response to changing match conditions (e.g., performance swings, resource imbalances, or strategic failures).

To address this gap, we draw on the General Aggression Model (GAM [4]), which offers a process-level, micro-level perspective on how aggression unfolds. GAM conceptualizes aggression as arising from the joint influence of situational inputs (e.g., provocation, frustration, environmental cues) and person factors (e.g., traits, attitudes, prior experiences) on an individual's internal state, including cognition, affect, and arousal. These internal routes are then translated, via appraisal and decision processes, into either impulsive or more deliberate behavioral responses. We adapt this input–route–appraisal structure to the context of online multiplayer matches, making the factors around toxic behaviors explicit at the level of concrete in-game events and moment-to-moment decision flows. In doing so, it complements prior work on discrete, static factors [36, 36–38]—such as player roles, trait dispositions, or saliency cues—by adding a process-oriented account of how internal perceptions, orientations, and intentions unfold around toxic events.

Our analysis finally unpacks the observed cognitive processes underlying aggressor behavior, identifying three interconnected layers—perception, orientation, and intention—that are shaped by both in-game emotional dynamics and latent cognition rooted in experiences beyond the game. In addition, our findings also align with prior evidence linking aggression levels to negative emotional valence [74], yet we also observe exceptions, such as harassment enacted for amusement, which underscore the complexity of emotion–cognition interactions within specific event–state flows and motivate a more nuanced, process-oriented model. Our current hierarchy does not fully specify how emotions and multi-level cognitive processes co-evolve during escalation and de-escalation. Future work can address this gap by turning existing emotion–cognition theories into explicit process models for toxicity. For example, Scherer [64]'s multi-level sequential check model conceptualizes appraisal as a structured process in which innate, learned, and deliberate evaluations proceed through checks such as novelty, goal relevance, implications for goals, coping potential, and normative significance. Likewise, Roseman and Smith [63] details how appraisals along dimensions like motive consistency and accountability, together with their intensity and certainty, combine to elicit specific emotions and action tendencies. Our hierarchy is broadly compatible with these process-oriented, multi-layer models: it situates players' ongoing evaluations within a layered structure in which moment-to-moment perceptions feed into more stable orientations toward teammates and opponents, which in turn shape short-term intentions to escalate, withdraw from, or repair toxic encounters. Building on existing models[63, 64], future work could therefore extend our observation into a more complete emotion–cognition process model of toxicity, specifying when and how specific appraisal patterns give rise to the diverse emotional profiles and behavioral pathways observed in our data.

Further, at the macro level, we identify two distinct toxicity trajectories: persistent and single-instance. We show how fragmented in-game events are linked by continuous internal states, such that earlier experiences can shape later decisions and sustain or dampen toxic behavior over time. This process view contributes a temporal, episode-based understanding of toxicity that goes beyond incident-level descriptions. In turn, it suggests a different design focus for intervention: instead of reacting only to isolated toxic acts, systems should monitor historical patterns to detect and defuse early signs of escalation, and provide mechanisms to repair players' internal states after incidents, thereby building a more preemptive line of defense against toxicity.

Taken together, our work contributes to a three-layer cognitive structure (perception, orientation, intention) that explains how in-game events and emotions jointly shape toxic behavior, and a

macro-level perspective that distinguishes persistent from single-instance toxicity trajectories, revealing how fragmented events are connected through continuous internal states and motivating process-oriented, preemptive intervention strategies.

6.2 Future Direction

By highlighting toxic behavior as a temporally embedded process rather than a static trait, our work suggests several directions for future real-time game analysis.

(1) More systematically link in-game events to changes in players' internal states, while accounting for out-of-game contexts. Case-based and longitudinal designs could combine fine-grained event logs with repeated self-reports or physiological measures to reconstruct how toxic episodes unfold over time, reveal how the same situation is appraised differently given players' goals, histories, and social backgrounds, and show how these person-by-context configurations shape trajectories of irritation, anger, or resignation.

(2) Map the bidirectional interplay between emotion and cognition during escalation by turning existing theories into an explicit process model. Our hierarchy is compatible with process-oriented, multi-layer emotion–cognition accounts, but our findings only sketch this linkage at a high level. As we mentioned above, future work could draw on appraisal-based emotion–cognition models [63, 64] to specify how emotional states and appraisals update one another over time and to formalize these feedback loops into a more complete process model.

(3) Trace how specific emotion–cognition configurations translate into concrete toxic actions, connecting micro-level psychological processes with observable behaviors. Rather than treating toxicity as a unitary outcome, future research can distinguish patterns such as hostile attribution with anger, contempt with moral disengagement, or frustration with perceived inefficacy, and examine how each pattern is tied to particular behavioral signatures like verbal abuse, griefing, or strategic withdrawal. By grounding behavioral categories in theoretically meaningful constellations of feelings and interpretations, this work can generate mechanism-based markers that are more informative for detecting, predicting, and tailoring interventions to different forms of toxicity.

Addressing these gaps may sharpen theoretical models and extend the practical window for preventive intervention by indicating where in an unfolding episode monitoring and just-in-time support are most likely to succeed.

6.3 Implication for Practice: Preventive Interventions and Designs for Toxic Behavior Mitigation in MOBA Games

Our proposed patterns no longer treat toxic behavior as a static property of information. Instead, they reveal that the generation of toxic behavior is a subtle and multi-layered process. The cognitive, emotional, and temporal mechanisms identified in our research form the foundation for a game toxic behavior evolution system. By targeting the evolution of these underlying factors, we can effectively intervene and disrupt the progression of toxicity. This enables a spectrum of design interventions that move beyond reactive moderation or post-hoc punishment, which establishes a preventive barrier before toxic behaviors occur. Based on this, we discuss how the prevention and de-escalation can be achieved by reshaping the interaction environment itself.

6.3.1 Early Team Formation and Impression Building. When players lack a shared sense of team identification, conflicts in MOBA environments tend to shift from collaborative problem-solving to individual attribution and blame [12, 40, 47]. This weakened team identification reduces the perceived social cost of initiating conflict, enabling negative internal states to be mobilized more readily into overt toxic behaviors [37, 42, 49]. Our findings suggest that early identity reinforcement may strengthen the commitment to collective outcomes and reduce hostility under stress. In some competitive MOBAs, the Ban & Pick is intended for hero selection and player communication [22]. However, we observed that, influenced by the toxic environment, players are often reluctant to communicate proactively. Future designs could consider introducing chatbot assistants to break the ice and facilitate the formation of team identity [67]. Furthermore, providing in-depth tactical communication support for macro team strategies (not just character selection) during the preparation phase may help players shift their attention from individual gains to shared team narratives, thereby providing a cognitive buffer against the escalation of in-game hostility.

6.3.2 Positive Information Dissemination and Friendly Communication Encouragement. Early positive interactions—such as offering assistance in critical situations—were found to rapidly establish immediate trust when perceived by other players, delaying the onset of hostility triggered by mistakes or pressure [12, 47, 82]. This buffering effect can be understood as a "compensation" regulatory mechanism: when first impressions included salient friendly signals, players tended to increase their tolerance toward teammates and were more inclined to forgive mistakes rather than assign blame. However, in the current gaming environment, players are more likely to focus on negative behaviors rather than such supportive friendly actions, similar to how assists or vision scores are often overlooked [47, 56, 62]. From a design perspective, amplifying the salience of early cooperative acts could strengthen this protective effect. For example, making invisible moments—such as a teammate covering you while you secure a resource—explicitly visible to players. Yet such visibility tools must be balanced; overemphasizing assistance could also amplify awareness of unhelpful teammates, potentially undermining team cohesion. Alternatively, systems could encourage micro-commitments to past friendly events during natural between-round or post-objective pauses—for instance, offering quick "commend for cooperation" prompts or encouraging players to acknowledge teammates after a successful cooperation. By reactivating earlier trust-establishing moments, these mechanisms may enhance the persistence of a positive team atmosphere under competitive stress (trust reinforcement [30]).

6.3.3 From Content Filtering to Contextual Intervention. A central implication of our findings is the need to move beyond static content filtering toward contextual intervention [65, 66, 77]. Toxicity assessments in competitive games go beyond isolated aggressive text; rather, they emerge from a multidimensional evaluation encompassing timing, frequency, scope, and perceived legitimacy. Many toxic behaviors originate from emotionally charged language, such as taunting and belittling. Some players express malice through

segmented typing or the use of homophones. These semantic expressions are highly contextualized. To address such toxic behavior, system design should prioritize detecting high-risk behavioral sequences over individual text. Such detection is also beneficial in mitigating persistent toxic behavior. Our analysis reveals that players' toxicity worsens through repeated exposure to the same antagonistic context, and that individual text filters fail to sever the interaction channels that fuel this persistence. Contextual detection helps identify persistent toxic behavior patterns and enables early intervention during their formation. These interventions need not be punitive; disrupting the escalation loop itself can meaningfully reduce affective contagion and persistent hostility. For example, employing structural circuit breakers—such as temporarily muting only in high-friction dyads [62, 77] or replacing provocative words with more neutral alternatives [61, 78] can interrupt this negative feedback cycle. Alternatively, gentle redirection can shift players' attention; for instance, the UI could offer an alternative interaction, such as a prompt to suggest the next team objective [42]. Through contextualized detection and by embedding alternative pathways at each decision point within the perception–orientation–intention hierarchy, systems can reduce behavioral transgressions and break escalation cycles without undermining essential collaborative communication.

6.3.4 Emotion Repair: Release Rather Than Simmer. Our study indicates that toxic behavior in competitive play can be amplified by uncontrolled emotional responses, with "emotional venting intent". This points to the value of integrating emotional self-regulation supports into system design as a preventive measure. Currently, most game designs lack dedicated channels for players to vent their emotions. Suppressing players' toxic behavior is not the end to mitigating toxic behavior; rather, it is essential to provide healthy outlets for players to process and release their emotions, or it may increase toxicity, *e.g.,* leading to disengagement. Systems can function as emotional release valves by redirecting venting into private channels, such as comment queuing [73], which allows temporarily toxic content to be placed in a pending queue, giving the sender an opportunity to decide whether to revise their expression before it is posted publicly. Interventions can also be adopted to help players shift their attention or conduct cognitive reappraisal. [15] Our results show that bystanders fail to react to or mitigate toxic behavior; system-mediated comfort may compensate. Prior work shows that social reassurance can buffer stress [14, 31], suggesting that timely comfort cues can aid emotional recovery in high-pressure settings. Context-triggered interventions might include AI-generated positively framed messages, emphasis on benefits, or supportive pings tailored to recent gameplay [5, 81]. By institutionalizing comfort as a system-level teammate function, games can help prevent negative emotional simmer.

7 Limitation and Future Work

Our study has several limitations. First, our small sample (18 players) was drawn solely from Mainland China servers. While our qualitative approach offers in-depth insight into cognitive and emotional processes during MOBA interpersonal conflict, it lacks statistical generalizability across game genres, regions, or cultures and focuses only on common forms of toxic behavior (excluding, for example, technical cheating, account boosting, or gendered, religious, or nationality-based discrimination). Given potential cross-cultural differences in norms, communication, and interpretations of "toxic" behaviors [25, 40, 41], future work should use larger, more diverse samples to test the external validity of our model. For unobserved instances, studies of gendered or racialized harassment could collect recordings of such incidents to capture in-game context and distinguish one-off occurrences from persistent toxicity. At the user level, researchers could invite aggressors to self-disclose (*e.g.,* , via surveys or interviews) to probe their subjective experiences and reconstruct perceived causal chains (*e.g.,* emotional venting vs. trait-driven behavior), thereby extending our framework to a broader range of cases. Second, our data is limited to aggressors' self-reports and retrospective accounts. We did not collect parallel data from corresponding victims, preventing systematic comparison between aggressors' interpretations and victims' experiences or perceived harm. Future work should include both perspectives from the same games and conduct cross-analysis to more directly reveal how toxic behaviors are interpreted differently across roles. Third, our use of retrospective think-aloud protocols, even with video playback, is vulnerable to memory bias, post-hoc rationalization, and self-presentation effects. Future work could use in-situ data capture and explicitly control for the observation effect. Fourth, our emotional analysis relied on self-reports visualized with journey maps, which capture subjective experience but may distort, like missing rapid in-game fluctuations. Future work could integrate real-time physiological measures (*e.g.,* HRV, GSR, facial analysis) [55, 71] synchronized with game events to provide a more granular view of emotional dynamics. Finally, our design implications remain conceptual. Developing and evaluating prototypes, *e.g.,* contextual intervention tools, through A/B testing or controlled studies is essential to assess their practical effectiveness in reducing toxicity and supporting collaboration.

8 Conclusion

This study examines how toxic behaviors emerge among *League of Legends* players by tracing aggressors' evolving cognition and emotion and identifying patterns in toxicity development. We show that aggressors judge the severity of their actions using multidimensional criteria, with cognition organized hierarchically: situation, event, and player perceptions shape an orientation layer (offensive, defensive, cooperative, indifferent), which then informs the intention layer that drives behavior. This hierarchy is influenced by both emotions and latent personal cognition. While toxicity generally escalates as emotions worsen, it can be an exception when causing distress is itself entertaining. Toxic behaviors manifest as persistent or single-instance acts, with single instances often intensified by short intervals and similar contexts. These findings highlight the need for both real-time/post-hoc interventions and preventive measures that block escalation at its onset. By clarifying how cognition and emotion evolve during the emergence of toxicity, our work provides a basis for preventive MOBA design interventions.

Acknowledgments

This project is supported by the Hong Kong SAR Research Grants Council's Theme-based Research Grant Scheme (Project No. T43-518/24-N). We are grateful to the anonymous reviewers for their valuable feedback and to our interview participants for their crucial contributions to advancing this research. We also sincerely thank Yulin Tian and other friends for the insightful perspectives they provided in the daily discussion.

References

[1] Sonam Adinolf and Selen Turkay. 2018. Toxic behaviors in Esports games: player perceptions and coping strategies. In *Proceedings of the 2018 Annual Symposium on computer-human interaction in play companion extended abstracts*. 365–372.

[2] Jesús C Aguerri, Mario Santisteban, and Fernando Miró-Llinares. 2023. The enemy hates best? Toxicity in league of legends and its content moderation implications. *European Journal on Criminal Policy and Research* 29, 3 (2023), 437–456.

[3] Sirwan Khalid Ahmed, Ribwar Arsalan Mohammed, Abdulqadir J Nashwan, Radhwan Hussein Ibrahim, Araz Qadir Abdalla, Barzan Mohammed M Ameen, and Renas Mohammed Khdhir. 2025. Using thematic analysis in qualitative research. *Journal of Medicine, Surgery, and Public Health* 6 (2025), 100198.

[4] Craig A Anderson and Brad J Bushman. 2002. Human aggression. *Annual review of psychology* 53, 1 (2002), 27–51.

[5] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. 2020. Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–30.

[6] 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*. 1–15.

[7] Nicole A Beres, Madison Klarkowski, and Regan L Mandryk. 2023. Playing with emotions: A systematic review examining emotions and emotion regulation in esports performance. *Proceedings of the ACM on Human-computer Interaction* 7, CHI PLAY (2023), 558–587.

[8] Thom Bongaards, Maurits Adriaanse, and Julian Frommel. 2024. Personalized Matchmaking Restrictions for Reduced Exposure to Toxicity: Preliminary Insights from an Interview Study. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play*. 31–36.

[9] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

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

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

[12] Alexandra Buchan and Jacqui Taylor. 2016. A qualitative exploration of factors affecting group cohesion and team play in multiplayer online battle arenas (mobas). *The Computer Games Journal* 5, 1 (2016), 65–89.

[13] Alessandro Canossa, Dmitry Salimov, Ahmad Azadvar, Casper Harteveld, and Georgios Yannakakis. 2021. For honor, for toxicity: Detecting toxic behavior through gameplay. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–29.

[14] Sheldon Cohen and Thomas A Wills. 1985. Stress, social support, and the buffering hypothesis. *Psychological bulletin* 98, 2 (1985), 310.

[15] Thomas F Denson, Michelle L Moulds, and Jessica R Grisham. 2012. The effects of analytical rumination, reappraisal, and distraction on anger experience. *Behavior therapy* 43, 2 (2012), 355–364.

[16] Na Du, Feng Zhou, Elizabeth M Pulver, Dawn M Tilbury, Lionel P Robert, Anuj K Pradhan, and X Jessie Yang. 2020. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation research part C: emerging technologies* 112 (2020), 78–87.

[17] Esports Insider. 2025. *Most popular esports in 2025: top 5 games right now.* https://esportsinsider.com/most-popular-esports-games-2025 Accessed September 9, 2025.

[18] Charles R Figley. 2012. *Encyclopedia of trauma: An interdisciplinary guide*. Sage Publications.

[19] Julian Frommel and Regan Mandryk. 2022. Effective Toxicity Prediction in Online Multiplayer Gaming: Four Obstacles to Making Approaches Usable. In *Mensch und Computer 2022-Workshopband*. Gesellschaft für Informatik eV, 10–18420.

[20] 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*. 1–12.

[21] Julian Frommel, Regan L Mandryk, and Madison Klarkowski. 2022. Challenges to Combating Toxicity and Harassment in Multiplayer Games: Involving the HCI Games Research Community. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. 263–265.

[22] Riot Games. n.d.. *How to Play.* https://www.leagueoflegends.com/en-gb/how-to-play/ Accessed September 9, 2025.

[23] Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Offensive language detection on video live streaming chat. In *Proceedings of the 28th international conference on computational linguistics*. 1936–1940.

[24] Nikos Giakoumoglou. 2025. *Game On, Rant On: Diary Insights of Toxic Triggers in League of Legends*. Master's thesis.

[25] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[26] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications.

[27] Mikko Halonen. 2024. The Use of Taboo Language in a Corpus of Chat Messages of Defence of the Ancients 2. (2024).

[28] Tharon Howard. 2014. Journey mapping: A brief overview. *Communication Design Quarterly Review* 2, 3 (2014), 10–13.

[29] Daniel Johnson, Lennart E Nacke, and Peta Wyeth. 2015. All about that base: differing player experiences in video game genres and the unique case of moba games. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 2265–2274.

[30] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546.

[31] Takefumi Kikusui, James T Winslow, and Yuji Mori. 2006. Social buffering: relief from stress and anxiety. *Philosophical Transactions of the Royal Society B: Biological Sciences* 361, 1476 (2006), 2215–2228.

[32] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* 30, 4 (2020), 1081–1102.

[33] Bastian Kordyaka, Sukran Karaosmanoglu, and Samuli Laato. 2025. Defining toxicity in multiplayer online games: A systematic literature review. *Computers in Human Behavior Reports* (2025), 100698.

[34] Bastian Kordyaka, Michael Klesel, and Katharina Jahn. 2019. Perpetrators in league of legends: scale development and validation of toxic behavior. (2019).

[35] Bastian Kordyaka, Samuli Laato, Juho Hamari, Tobias Scholz, and Björn Niehaves. 2023. What drives gamer toxicity? Essays from players. In *GamiFIN Conference*. CEUR-WS.

[36] Bastian Kordyaka, Samuli Laato, Katharina Jahn, Juho Hamari, and Bjoern Niehaves. 2023. The cycle of toxicity: Exploring relationships between personality and player roles in toxic behavior in multiplayer online battle arena games. *Proceedings of the ACM on human-computer interaction* 7, CHI PLAY (2023), 611–641.

[37] Bastian Kordyaka, Samuli Laato, and Bjoern Niehaves. 2024. A toxic triad: Aggression, anger and authoritarianism-A study with multiplayer online battle arena game players. In *International GamiFIN Conference*. CEUR Workshop Proceedings.

[38] Bastian Kordyaka, Samuli Laato, Sebastian Weber, and Gerhard Klassen. 2023. The Saliency of Dispositions: Personality Traits, Anger. and Aggression as Antecedents of Toxicity in Multiplayer Online Battle Arena Games. (2023).

[39] Bastian Kordyaka, Luisa Pumplun, Marlies Brunnhofer, Bjoern Kruse, and Samuli Laato. 2023. Gender disparities in esports–An explanatory mixed-methods approach. *Computers in Human Behavior* 149 (2023), 107956.

[40] 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*. 81–92.

[41] Yubo Kou. 2021. Punishment and its discontents: An analysis of permanent ban in an online game community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–21.

[42] Yubo Kou and Xinning Gui. 2020. Emotion regulation in esports gaming: A qualitative study of league of legends. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[43] Rachel Kowert. 2020. Dark participation in games. *Frontiers in Psychology* 11 (2020), 598947.

[44] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3739–3748.

[45] 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 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[46] Samuli Laato, Bastian Kordyaka, Velvet Spors, and Juho Hamari. 2024. How StarCraft II Players Cope with Toxicity: Insights from Player Interviews. In *International Conference on Human-Computer Interaction*. Springer, 203–219.

[47] Juhoon Lee, Seoyoung Kim, Yeon Su Park, Juho Kim, Jeong-woo Jang, and Joseph Seering. 2025. Less Talk, More Trust: Understanding Players' In-game Assessment of Communication Processes in League of Legends. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[48] 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*. 1–21.

[49] Charles K Monge and TC O'brien. 2022. Effects of individual toxic behavior on team performance in League of Legends. *Media Psychology* 25, 1 (2022), 82–105.

[50] Marçal Mora-Cantallops and Miguel-Ángel Sicilia. 2018. MOBA games: A literature review. *Entertainment computing* 26 (2018), 128–138.

[51] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security* 76 (2018), 197–213.

[52] Joaquim AM Neto, Kazuki M Yokoyama, and Karin Becker. 2017. Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the international conference on web intelligence*. 26–33.

[53] Newzoo. 2025. *League of Legends: Game profile and market data.* https://web.archive.org/web/20240920092715/https://newzoo.com/games/league-of-legends Accessed: 2025-09-08, Archived from Newzoo.com.

[54] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847.

[55] Varsha Kiran Patil, Vijaya R Pawar, Shreiya Randive, Rutika Rajesh Bankar, Dhanashree Yende, and Aditya Kiran Patil. 2023. From face detection to emotion recognition on the framework of Raspberry pi and galvanic skin response sensor for visual and physiological biosignals. *Journal of Electrical Systems and Information Technology* 10, 1 (2023), 24.

[56] Vicente Peñarroja, Virginia Orengo, Ana Zornoza, and Ana Hernández. 2013. The effects of virtuality level on task-related collaborative behaviors: The mediating role of team trust. *Computers in Human Behavior* 29, 3 (2013), 967–974.

[57] Matthew D Pickard, Catherine A Roster, and Yixing Chen. 2016. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior* 65 (2016), 23–30.

[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. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 1–26.

[59] Michal Prokop, Ladislav Pilař, and Ivana Tichá. 2020. Impact of think-aloud on eye-tracking: A comparison of concurrent and retrospective think-aloud for research on decision-making in the game environment. *Sensors* 20, 10 (2020), 2750.

[60] Kevin Proudfoot. 2023. Inductive/deductive hybrid thematic analysis in mixed methods research. *Journal of mixed methods research* 17, 3 (2023), 308–326.

[61] N Prudhvish, G Nagarajan, U Bharath Kumar, B Harsha Vardhan, and L Tharun Kumar. 2024. DeTox: A WebApp for Toxic Comment Detection and Moderation. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*. IEEE, 1–5.

[62] 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. *Proceedings of the ACM on human-computer interaction* 6, CHI PLAY (2022), 1–27.

[63] Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* (2001), 3–19.

[64] Klaus R Scherer. 1999. Appraisal theory. (1999).

[65] Adrien Schurger-Foy, Rafal Dariusz Kocielnik, Caglar Gulcehre, and R Michael Alvarez. 2025. Context-Aware Toxicity Detection in Multiplayer Games: Integrating Domain-Adaptive Pretraining and Match Metadata. *arXiv preprint arXiv:2504.01534* (2025).

[66] Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing* 490 (2022), 312–318.

[67] Donghoon Shin, Soomin Kim, Ruoxi Shang, Joonhwan Lee, and Gary Hsieh. 2023. IntroBot: Exploring the use of chatbot-assisted familiarization in online collaborative groups. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[68] Statista. 2025. *League of Legends - Statistics & Facts.* https://web.archive.org/web/20240920093021/https://www.statista.com/topics/4266/league-of-legends/#topicOverview Accessed: 2025-09-08, Archived from Statista.com.

[69] Natalia Stepanova, Wesley Muthemba, Ross Todrzak, Michael Cross, Nicholas Ames, and John Raiti. 2021. Natural language processing and sentiment analysis for verbal aggression detection; a solution for cyberbullying during live video gaming. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*. 117–118.

[70] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[71] Mikel Val-Calvo, José Ramón Álvarez-Sánchez, Jose Manuel Ferrández-Vicente, Alejandro Díaz-Morcillo, and Eduardo Fernández-Jover. 2020. Real-time multimodal estimation of dynamically evoked emotions using EEG, heart rate and galvanic skin response. *International journal of neural systems* 30, 04 (2020), 2050013.

[72] Maaike J Van den Haak and Menno DT De Jong. 2003. Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. In *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.* IEEE, 3–pp.

[73] AKRITI VERMA. 2024. Encouraging Self-Reflection in Online Conversations: A Comment Queuing Approach to Mitigating Toxicity and Enhancing Emotional Regulation. (2024).

[74] Edelyn Verona and Konrad Bresin. 2015. Aggression proneness: Transdiagnostic processes involving negative valence and cognitive systems. *International Journal of Psychophysiology* 98, 2 (2015), 321–329.

[75] Michel Wijkstra. 2024. Fighting Toxicity Through Positive and Preventative Intervention. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play*. 450–453.

[76] 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*. 3–9.

[77] 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. *Proceedings of the ACM on human-computer interaction* 8, CHI PLAY (2024), 1–32.

[78] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. 2021. RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–26.

[79] 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 Transactions on Computer-Human Interaction* 30, 6 (2023), 1–36.

[80] Mengyun Yao, Yuhong Zhou, Jiayu Li, and Xuemei Gao. 2019. Violent video games exposure and aggression: The role of moral disengagement, anger, hostility, and disinhibition. *Aggressive behavior* 45, 6 (2019), 662–670.

[81] Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of AI in peer support for young people: A study of preferences for human-and AI-generated responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[82] Yilei Zeng, Anna Sapienza, and Emilio Ferrara. 2019. The influence of social ties on performance in team-based online games. *IEEE Transactions on Games* 13, 4 (2019), 358–367.

[83] Zinan Zhang, Sam Moradzadeh, Andrew Woan, and Yubo Kou. 2024. Toxicity by Game Design: How Players Perceive the Influence of Game Design on Toxicity. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY (2024), 1–31.

[84] Ágnes Zsila, Reza Shabahang, Mara S Aruguete, and Gábor Orosz. 2022. Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences. *Aggressive Behavior* 48, 3 (2022), 356–364.

A Toxic Behavior Taxonomy

Toxic behavior	Definition	Observed incidents	Participants involved (n)
Insulting and verbal abuse	Saying foul things to the other player, calling them names or insinuating that there is something wrong with them.	22	13
Provoking and taunting	Trying to make the other player angry, e.g., taunting in front of the enemy, or provoking/insulting others verbally.	25	13
Whining and presenting emotionally	Complaining about other players or presenting emotionally fueled.	35	16
Spamming	Repeatedly engaging in an action such as sending the same message.	6	2
Griefing	Irritating and/or harassing other players by using the game in unintended ways, such as intentional feeding or AFK.	4	3
Mediocritizing	Gameplay actions that do not maximize the winning chance, acting passively or not putting in effort.	7	5
Lack of adherence to social norms	Deviating from the established behavioral conventions in a way that throws the opponent off.	4	2
Hostage holding	Purposely keeping other players in an unpleasant situation, such as preventing the game from ending.	2	2
Cheating	Gaining an unfair advantage; includes scripting, smurfing, and rank boosting.	0	0
Sexual harassment / Hate speech / Personal threats	Insults or comments based on gender, religion, race, nationality, personal letters, and threats against individuals.	0	0

Table 3: Toxic behaviors and definitions synthesized from previous literature, together with descriptive counts in our dataset.