How To Tame a Toxic Player? A Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games

MICHEL WIJKSTRA, Utrecht University, Netherlands
KATJA ROGERS, University of Amsterdam, Netherlands
REGAN L. MANDRYK, University of Victoria, Canada
REMCO C. VELTKAMP, Utrecht University, Netherlands
JULIAN FROMMEL, Utrecht University, Netherlands

Toxic behavior is known to cause harm in online games. Players regularly experience negative, hateful, or inappropriate behavior. Interventions, such as banning players or chat message filtering, can help combat toxicity but are not widely available or even comprehensively studied regarding their approaches and evaluations. We conducted a systematic literature review that provides insights into the current state of interventions literature, outlining their strengths and shortcomings. We identified 36 interventions and qualitatively analyzed their approaches. We describe the types of toxicity being addressed, the entities through which they act, the methods used by intervention systems, and how they are evaluated. Our results provide guidance for future interventions, outlining a design space based on known systems. Furthermore, our findings highlight gaps in the literature, e.g., a sparsity of empirical evaluations, and underexplored areas in the design space, enabling researchers to explore novel directions for future interventions.

CCS Concepts: • **Applied computing** → **Computer games**; • **Software and its engineering** → **Interactive games**; • **General and reference** → **Surveys and overviews**; • **Human-centered computing** → **Collaborative and social computing**.

Additional Key Words and Phrases: toxicity, interventions, systematic literature review, online games

ACM Reference Format:
Michel Wijkstra, Katja Rogers, Regan L. Mandryk, Remco C. Veltkamp, and Julian Frommel. 2024. How To Tame a Toxic Player? A Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY, Article 315 (October 2024), 32 pages. https://doi.org/10.1145/3677080

1 Introduction

Toxicity is a problem that plagues players and developers of most multiplayer games [55, 56]. The problem first found its way into academic literature with work by Dibbell in 1994 [19]. The Anti-Defamation League presented a critical report in 2022 [56], which revealed that five out of six adults (86%) have experienced harassment in online play. This is problematic at a large scale, as toxic actions disrupt the players' enjoyment and performance and can result in lasting harm [76].

Authors' Contact Information: Michel Wijkstra, m.wijkstra@uu.nl, Utrecht University, Utrecht, Netherlands; Katja Rogers, k.s.rogers@uva.nl, University of Amsterdam, Amsterdam, Noord-Holland, Netherlands; Regan L. Mandryk, reganmandryk@uvic.ca, University of Victoria, Victoria, British Columbia, Canada; Remco C. Veltkamp, Utrecht University, Utrecht, Netherlands, r.c.veltkamp@uu.nl; Julian Frommel, Utrecht University, Utrecht, Netherlands, j.frommel@uu.nl.

ACM 2573-0142/2024/10-ART315
https://doi.org/10.1145/3677080

While game developers, researchers, and players recognize the problem and have started combating toxicity, it has not been solved but has instead grown in severity.

One approach for dealing with toxic behavior and its effects in multiplayer games are *intervention* systems. We define an intervention system as any digital system component that helps combat toxicity or its effects. Many of these approaches have been common in commercial games for a long time, such as reporting a player in a session, banning them from the server, and automated systems to monitor player behavior (e.g., detection of harassment, hate speech, or disruptive behavior) [63, 70, 73, 89]. In addition to approaches in commercial games, there is also intervention-focused work in academic research that helps combat toxic behavior. For example, prior work has proposed new interventions (e.g., [13, 58, 72]) or analyzed the efficacy of existing approaches (e.g., [11, 26, 96]). There is evidence for the value of interventions in non-gaming contexts [25], commercial games [69, 81], and games research [76]. Despite the interest in toxicity interventions from diverse stakeholders who hold in common the desire to prevent and mitigate harm, and despite the awareness that addressing toxicity is a priority for game companies and communities alike, efforts to combat toxicity have not yet seen widespread success. The companies, researchers, communities, and players who aim to combat toxicity would benefit from a shared strategy and a culture of cooperation and information exchange—echoes of which are seen in organizations like the Fair Play Alliance[1]—to work toward a common goal. The *problem is that designers, researchers, and policymakers have little guidance about which approaches to implement in a game, how to assess the effectiveness of interventions, or which directions of intervention development hold the most promise.* While there is valuable literature about toxicity intervention, there is not yet much synthesis of this work. In this paper, we build on our work-in-progress [103] and further analyze the same dataset to provide an in-depth overview of research about interventions for game toxicity, to answer several novel research questions (see Section 3), and to create a framework of intervention approaches, and to expose promising innovation directions.

We have conducted a systematic review of prior work on intervention systems for toxic behaviors, based on a protocol aligned with the PRISMA-P standard [71, 86] and pre-registered on OSF. Using a systematic database search, we collected literature on papers that discussed addressing toxicity through intervention systems. After abstract screening, full-text screening, and several rounds of discussions, we identified 36 intervention systems in 30 papers. Through thematic analyses, we mapped different approaches along research questions and into overarching themes guiding a design space for interventions and their evaluation approaches.

In our analysis, we defined themes for types of toxic behavior (*general* and *targeted toxic behavior*), the sites through which intervention systems act (*player* and *platform*), and methods used by intervention systems (*empowering*, *supporting*, *priming*, *sanctioning*, and *detecting*), mapping a design space for intervention systems. We further found that most systems act only *after* toxicity has occurred, highlighting gaps in preventative approaches, while also focusing on proposing *novel* instead of analyzing *existing* approaches. When analyzing evaluation practices, we map different approaches to evaluation, while also highlighting gaps in evaluation with players and in commercial settings.

Overall, these results show that there are gaps in the design space of toxicity interventions and their evaluation designs. Through this systematic literature review, we provide insights into the state of toxicity intervention literature and provide guidance that can inform future intervention designs by enabling researchers to explore novel directions.

[1] https://fairplayalliance.org/

2 Background

2.1 Toxicity

Toxic behavior is a common occurrence in online spaces. While one might expect that the playful nature of games makes them an exception, nothing is less true. Many, if not all, online games suffer from toxicity in their communities. Toxicity is an umbrella term used to describe a collection of negative behaviors [97]. It is difficult to precisely define what toxicity is, as the context of the behaviors, and even the offline environment of the community members play a large role in its perception [48]. Furthermore, where a certain type of behavior could be considered acceptable in a group of adults, the social rules are different for other groups, such as when children are in the context. Lastly, the meaning of toxicity depends on the context of the game. Certain actions could be considered normal or even an important gameplay element in one game, while being considered negative in another. For example, it is a core gameplay element in the game Rust [91] to kill other players, raid their houses, and destroy their belongings. While essential in this game, such actions could be considered toxic in other games. In practice, we consider behavior to be toxic if it violates the rules and social norms that apply in the game and community it occurs in [6].

Both players and game developers suffer from the negative impacts of toxic behavior. Players experience loss of enjoyment [97], lower in-game performance [7, 47, 107], distress [26], a sense of losing control [76], and on a larger scale diminished quality of community feel due to toxicity enabling more toxicity [49]. Game developers are not excluded from the negative effects of toxicity. It harms user retention [34], which is crucial in online games. Further, having a toxic community creates a negative association with the game, making it harder to attract new users [40, 65]. The online disinhibition effect is often cited as a strong driver of toxic behavior in online spaces [92]. This effect causes a lack of restraint when communicating in online spaces as compared to communicating face-to-face. Combined with the fast-paced nature of online video games, competitiveness and lack of consequences result in a high probability for repeated toxic behavior [47]. Prior work by Beres et al. [6] demonstrated that players with higher toxic online disinhibition and moral disengagement perceive behavior as less toxic. This highlights that subjectivity complicates this problem because different types of toxicity may be perceived as differently severe and thus potentially also require different interventions. In our study, we examine which types of toxic behaviors intervention systems attempt to combat (RQ1).

2.2 Intervention Systems

Intervention systems are an important part of combating harm, e.g., by helping detect, sanction, and in some cases prevent toxicity. More broadly, outside of an HCI & Games context, intervention systems are often understood as prevention strategies. For example, they have been studied and applied against undesirable human behavior ranging from broad social and emotional problems to specific issues like substance abuse and crime [3, 38, 66]. A large body of research focuses on education systems in the United States of America. This research has identified the need for well-timed and continued presence of interventions [66], the need for behavioral change, and the forming of positive social relationships [3] as important elements in successful strategies. Similarly, intervention systems are a prominent topic in research on online content moderation (often outside of game contexts), i.e., mechanisms to facilitate cooperation and prevent abuse [36]. Due to the differences in culture and dynamics of gameplay affecting how interventions will affect player experience, interventions are different in the games context, highlighting the need to specifically investigate research on interventions in this context. However, non-game moderation research provides useful lenses for our analysis. Joseph Seering [82] discusses two perspectives in moderation research: That of the platform and policies, and that of the community. We are similarly

interested in understanding through which vector (we call these sites) game toxicity interventions act (RQ2).

2.3 Moderation

The existence of different moderation methods has been well documented in the context of content moderation with trade-offs in moderation design [44]. This aligns with work by Seering et al. [83] who explain, based on deterrence theory, that multiple different approaches to moderation (e.g., chat modes restricting the ability to post certain content vs reactive bans) can reduce the spread of different types of behavior. We similarly study which methods are used by interventions in games (RQ3). Seering et al. [83] further differentiate two styles of moderation tools: proactive and reactive systems. Proactive systems prevent certain behaviors while reactive systems apply a punishment after a certain behavior occurs. In RQ4, we similarly explore this, investigating if intervention systems act before or after harm. Yen et al. [108] identified a collection of works that investigated the influences of sanctions on community users and found that they often highlight a flaw in current reactive approaches. These approaches appear to be ineffective at encouraging prosocial behavior. This is problematic, because real world intervention studies suggest that this is essential in an effective system. In the context of online harm, Feinberg and Robey [23] and Tanrikulu [95] published guidelines and a literature review of prevention strategies for cyberbullying. While cyberbullying is related to toxicity in online video games, most cyberbullying research does not specifically focus on games and their specific dynamics. In the broader context of cyberbullying, interventions often focus on preventing active, intentional bullying [95], e.g., training programs including awareness and empathy guided by the theory of planned behavior [2, 16]. Such directed prevention strategies may be valuable for directed toxicity, but not necessarily effective for other types of game toxicity, e.g., flaming due to frustration. In this paper, we review interventions in the context of game-based toxicity to provide a broader overview of the approaches.

By designing systems or game elements that actively prevent toxicity or by creating reactive systems that monitor behavior and enforce rules, game developers can prevent harm and exposure. Many intervention methods have been proposed, addressing different types of toxic behaviors and using multiple methods and approaches to intervene. For example, Blackburn and Kwak [8] evaluate toxicity prediction, i.e., predicting if messages are toxic, which can be used to apply sanctions to misbehaving players. Another possible approach is demonstrated by Reid et al. [76], who suggest supporting the victim instead of punishing the perpetrator. Work by Kou and Gui [52] investigates reporting systems, which is a common way of addressing toxicity. These works demonstrate that many approaches currently exist within academic literature.

In summary, many elements of moderation and intervention have been researched in different fields. Further, various intervention approaches and evaluation strategies exist in academic literature. However, no work has currently explored interventions for online games at large, analyzing their approaches, design elements, and evaluation practices.

3 Research Aims

With our work, we assess the current state of the literature on intervention methods for toxic behaviors in online video games. We aimed to answer the following research questions:

- **RQ 1**: Which types of toxic behaviors are interventions aiming to combat?
- **RQ 2**: What sites are researchers using to combat toxicity?
- **RQ 3**: What methods are used by intervention systems?
- **RQ 4**: Are interventions applied before or after harm from toxic behavior occurs?
- **RQ 5**: How are researchers evaluating the effectiveness of their intervention systems?

- **RQ 5.1**: What metric do authors use to evaluate their interventions?
- **RQ 5.2**: Do authors evaluate existing or novel interventions?
- **RQ 5.3**: Do authors evaluate their interventions in realistic settings?

4 Methods

We conducted a systematic literature review and designed a review protocol based on the PRISMA-P guidelines [86] using keyword definitions and anchor papers to guide a database search in abstracts and titles. The results of this database search were de-duplicated, resulting in 1176 unique papers that were screened for inclusion and 1146 papers that were subsequently removed. The remaining 30 papers were then coded according to their characteristics. This review protocol was pre-registered on OSF and is described below. Figure 1 displays the workflow used to gather the studies in our study.

4.1 Toxicity Scope, Keywords, and Anchor Papers

While defining the scope of our study, we followed the umbrella term definition by Türkay et al. [97], capturing various forms of negative behaviours exhibited by players in online environments. We included the following behaviors while defining our search query: harassment, abuse, hate speech, insulting, griefing, trolling, offending/offensive behavior, inappropriate behavior, dark participation, and abusive behavior. This builds on prior work that investigated different aspects of toxic and harmful behaviors [6, 47, 53, 97] to provide a more expansive coverage of the term. In contrast, we left out behaviors such as cheating and botting, as well as associated interventions, such as anti-cheat systems. While those behaviors can be toxic [74], they are often not intentionally harmful to users. For example, a reason for cheating could be to more quickly progress the players' own account or character, with the negative experience for others being a side effect. We used this scope to guide our search to collect a broad sample of toxicity research covering different aspects of behavior that harm other players and accordingly used them to define the keywords.

The keywords were selected to cover the behaviors we defined as our scope earlier in this section, and were modified to be searchable in their different tenses and variations by applying wildcards. These wildcards allowed us to capture variations of common words used to describe toxicity, e.g., `harass*` to capture the terms harass, harassed, harassment, and harasser. We also added keywords to limit the scope to online multiplayer games. With these keywords, we defined the following database search query: `(toxic* OR harass* OR hate* OR insult* OR grief* OR trol* OR offen* OR inappropriate OR "Dark Participation" OR abus* OR flaming) AND ("multiplayer game" OR "multiplayer games" OR "multiplayer gaming" OR "online game" OR "online games" OR "online gaming" OR "online play" OR esports OR "e-sports" OR "competitive game" OR "competitive gaming" OR "competitive games" OR "video games" OR "video game" OR "video gaming" OR MMO OR MOBA OR FPS)`. This query was adapted to work for the different databases and applied based on abstracts and titles (full queries are available in the pre-registration document).

Further, we selected a set of 10 papers that matched our defined scope during our initial experimentation with the search terms. These papers were used to verify the completeness of our search results before the application of our inclusion criteria. We made a conscious effort to create a diverse selection of works that aim to combat toxicity. Eight of these matched our selection criteria and acted as anchor papers for the inclusion: Reid et al. [76], Kou and Gui [52], Canossa et al. [13], Murnion et al. [64], Märtens et al. [65], Kou [50], Blackburn and Kwak [8], and Kaiser and Feng [45]. We selected these papers because they describe different approaches and goals to combat toxicity, covering different authors, fields, databases, and publication years. This was necessary to ensure broad coverage of the corpus, as we would later feed these papers to our ML assisted screening

software. Two papers about toxicity were selected to test if the query reached relevant toxicity literature: Kowert [53] and Kordyaka et al. [47]. All anchor papers should appear in the database search but these latter two papers would ultimately be excluded from the review. Thus, they served as exclusion anchor papers, as they match the topic because they describe fundamentals of toxicity but do not propose an intervention system. We used these papers to test the database query and to seed the active learning model used in the abstract screening phase.

4.2 Database Search

We selected four electronic libraries containing HCI & Games research: ACM Digital Library: The ACM Guide to Computing Literature (the full collection), IEEE Xplore, Scopus, and Web of Science. We limited our search to a date range from 1990 up to and including 2022. This date range was selected as it goes beyond our first known instance of toxicity research [19], up to the point when our data collection started. These collections returned an initial set of 1906 records. We imported this collection into the Zotero Reference Manager, which identified 730 duplicates that we excluded before we started the screening process. To test the query and comprehensiveness of the database search, we verified that our results included the 10 anchor papers described previously; all anchor papers were present in our collection.

We also observed that the ACM Digital Library provided us with a result that did not meet our search criteria. While our search query was limited to the abstract and title of the papers, in this specific case, our query would be matched with a small sample of the work's full text. We were unable to detect any such discrepancies for the other databases. This discrepancy concerned a single item by Tally et al. [94]: it did not fully match our toxicity keywords in the abstract or title, yet was still put forward by the ACM Digital Library search engine. We initially discussed including this paper. However, we removed it during full-text screening, as it did not meet the inclusion criteria (as defined in Section 4.3).

4.3 Abstract Screening

After de-duplication, we had 1176 papers left that were eligible for abstract screening. For this initial screening phase, we defined inclusion and exclusion criteria based on what was relevant for our review, i.e., any game-related work focusing on toxicity and describing a digital system component, also allowing works that focus on the broader ecosystem around games (e.g., publisher website, Discord community, or Steam community). We excluded works that describe offline or non-digital games, do not include a digital component, or do not explicitly focus on toxic behavior (as defined by our keywords). When we were unable to make a decision based on the abstract alone, we verified eligibility by screening the full text. Our full inclusion criteria were as follows:

(1) We consider any game-related work focusing on toxicity. As toxicity is generally used as an umbrella term, we include a variety of behaviors that has been considered part of toxicity such as, but not limited to: harassment, abuse, hate speech, insulting, griefing, trolling, offending/offensive behavior, inappropriate behavior, dark participation and abusive behavior.
(2) We include works that describe user-sourced toxicity prevention (e.g. clever or unintended use of in-game systems)
(3) The work has to describe a digital system component, including but not limited to:
   (a) Tools
   (b) Algorithms
   (c) AI models
   (d) Design approaches
   (e) We allow human-in-loop systems, e.g., moderators, tribunal (League of Legends)

(f) We include text-based components, e.g., Terms of Service prompts, loading screen messages

(4) In addition to components integrated directly in games, we also include components integrated in the broader ecosystem of games, e.g., publisher website or Discord

(a) We do so as communities often form outside of the main game application.

Our selection process was further guided by the following exclusion criteria:

(1) Work describes offline games
(2) Work describes non-digital games
(3) Work does not include a digital component
(4) Work does not have toxic behavior (as defined by our terms) or the addressing of such behaviors as its main subject.

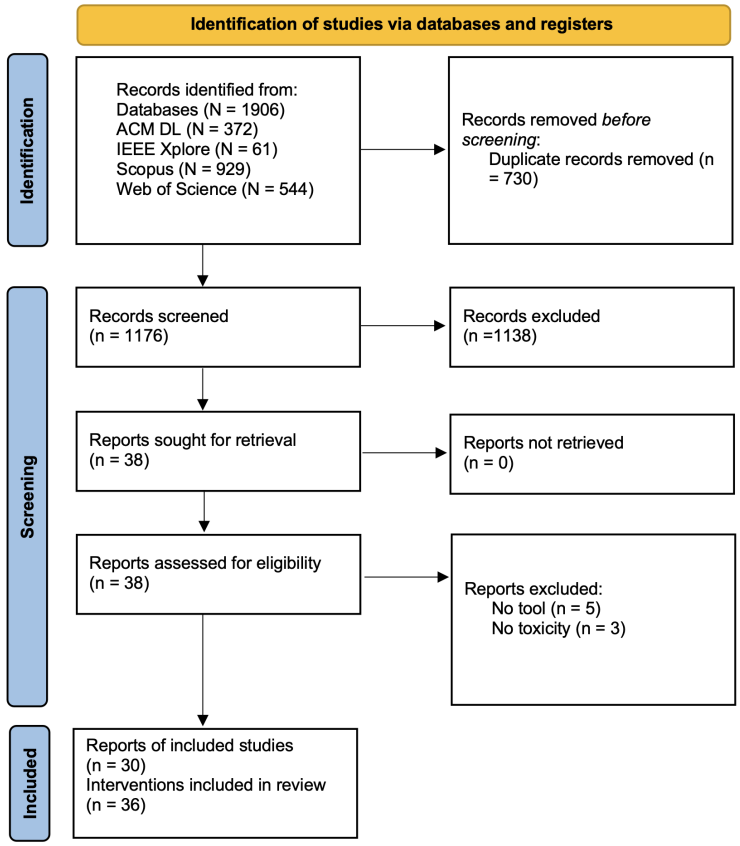

Fig. 1. Process of the review's search and selection stages, with results across the different steps.

We applied tool-assisted screening with active learning techniques using ASReview [99], which is an active learning tool aimed at helping authors throughout the title and abstract screening phase of systematic literature reviews. This approach reorders the set of items to review based on relevance, thus prioritizing work that is more likely relevant. In combination with carefully chosen stopping criteria, it can therefore reduce the number of papers to review while ensuring that the likelihood of missing relevant papers is low. Active learning approaches benefit from being initialized with labeled examples. For this, we used our anchor papers and an additional set of 10

random papers from our initial set of search results. The first and the last author both screened the abstracts of these random papers independently before they met and discussed their results. Both authors agreed on all papers regarding if they should be included. Based on this, we had 20 labeled papers (containing 8 relevant papers and 12 irrelevant papers), which were imported into ASReview to guide its ordering process.

The first author then screened the remaining abstracts until the previously defined stopping criteria were met. Our pre-defined stopping criteria were data-driven and stated that we stopped screening when a pre-defined number of papers were all excluded without a single relevant paper. We defined this number at 10% of the dataset (i.e., 118 papers). This threshold was reached after screening 488 abstracts, at which point we stopped the screening process. Based on benchmarks performed by the developers of ASReview, we can on average expect to reach 95% recall after screening at best 8% and at worst 36% of the collection size [24, 99]. As we processed 42% of our initial collection we are confident that we extracted the significant majority, if not all relevant works while keeping the time requirement for the screening process manageable. The resulting collection consisted of 38 papers.

4.4 Full-text Screening

The 38 papers selected in the abstract screening were then reviewed in full-text screening. During this phase, we verified eligibility once more, resulting in 8 more papers being excluded because they did not meet the inclusion/exclusion criteria. Common exclusion reasons included: 1) not providing a concrete tool, e.g., [35, 54] and 2) not focusing on toxicity, e.g., [37, 41]. After this, we had 30 relevant papers in our data set. From these papers, we extracted information using a pre-defined data extraction sheet. This extraction sheet consisted of 36 fields and questions such as *"Paper Title"*, *"Publication Year"*, and *"What method is used to reduce toxicity - Article Definition"*. Through the use of this sheet, we identified separate intervention approaches in each paper (e.g., seven interventions in [76]). With this process, we identified 36 unique intervention systems that we further analyzed. Figure 1 displays the entire process and results across the steps.

4.5 Analysis

After the full-text screening of all selected papers and finalizing the data extraction, we conducted a thematic analysis informed by the umbrella definition by Braun and Clarke [9] and further guided by Byrne [12]. The goal of this thematic analysis was to assess the current state of literature on intervention methods for toxic behaviors in online video games. The first author was already sufficiently familiar with the data after having read the papers. Throughout the reading, they iteratively and inductively coded the information about the interventions according to the data extraction sheet using semantic and latent codes. For each column in the data extraction sheet, a section of text was extracted as a semantic code (most commonly from the introduction, method, results, or conclusion sections). This was placed in the *"article definition"* field. The first author then interpreted this text and formulated a latent code that was more uniform with the rest of the papers, creating a first basic coding. This code was placed in the *"interpreted definition"* field for each question in the extraction sheet. For example, for the question ' *'What method is used to reduce toxicity"* the data extraction sheet featured *"To help protect its younger users by providing a family-friendly experience, [the publisher] needed a way to be able to filter out vulgarity in real time from all the instances where users are given the opportunity to submit text"* [18]. This was labelled with the code *"Chat Filtering"*.

After initial coding, the first and last author met to discuss the coding. This resulted in a second iteration of the coding approach, where we revised columns in the extraction sheet and defined a set of deductive codes for some of the columns. These deductive codes were required to answer

some of the research questions. For RQ4 (*"Are Interventions Applied Before or After Harm From Toxic Behavior Occurs?"*), we coded interventions as approaches based on when they are applied. We coded approaches as *after* toxicity if the intervention was applied after toxicity had already affected another player. We categorized interventions as *before* if they worked without a specific instance of toxicity affecting other players. An example of this would be chat filters [18], which are applied after a player makes a toxic remark, but *before* another player is affected by this. Importantly, we coded ambiguous approaches as *after* even if they had proactive components but relied on toxicity having occurred (e.g., blocking other players will prevent further exposure but only after the toxicity has already affected the targeted player). We make the distinction between before and after toxicity because acting before toxicity occurs has the benefit of preventing the occurrence of harm and lowering the chance of repeated negative behavior or acts of retribution. For RQ 5.2 (*" Do Authors Evaluate Existing or Novel Interventions?"*), we assessed the prevalence of research that provides new technical systems in comparison to understanding commonly used approaches. We coded the approaches as either *existing*—if the intervention already existed and was evaluated in the paper, e.g., [50, 96]—or as *novel*—if the intervention was proposed as a novel system as part of the paper, e.g, [29, 90]. Lastly, for RQ 5.3 (*"Do Authors Evaluate Their Interventions in Realistic Settings?"*), we coded approaches as *evaluated with players* if they were evaluated with player feedback (e.g., in a user study, with forum comments, or through voluntary player reports [10, 26]). We coded them as *not evaluated with players* if no such evaluation was performed. This is often the case for machine learning-based intervention systems, which are generally evaluated through statistical analysis of classification accuracy, e.g., [17, 72]. Furthermore, we coded interventions as *based on commercial settings* if they were created with or applied to data from a commercial game or platform, e.g., a League of Legends dataset [67]. We applied the code *not based on commercial settings* if the intervention method has not been applied to such a setting, e.g., instead, it was applied in a custom game designed for the experiment [29].

After this second round of coding, the first and last authors met several times to discuss codes and overarching commonalities and generate an initial set of themes. This involved affinity mapping in-person and with digital tools, as well as a discussion of the data and generated themes. In our analysis, we also allowed codes (and thus intervention systems) to contribute to multiple themes, e.g., a system focusing both on player mood and reducing exposure [76] was added to both those themes. Then, the remaining authors reviewed the themes with subsequent meetings to discuss them. Through this, we refined the groupings and defined the final themes. We generated separate sets of themes for RQs 1–3 and RQ 5.1. Based on these, the themes were developed as groupings for different types, sites, methods, and metrics relating to the toxicity interventions. These all aligned with our goal of defining a framework for toxicity interventions. Finally, through analysis of the theme-based groupings, we answered the research questions.

5 Results

In this section, we present our findings. Table 1 provides descriptions of each of the intervention systems in our study. This table also acts as a reference guide for specific interventions using the intervention IDs, as seen in Tables 2-5.

5.1 RQ 1: Which Types of Toxic Behaviors Are Interventions Aiming to Combat?

Within the umbrella term that defines toxicity, we have categorized two overarching types of toxicity. While assigning a severity level to a certain behavior is challenging, there are differences. The first theme we defined is general toxic behavior. The second theme consists of *targeted attacks* against an individual, which are more harmful and never acceptable.

5.1.1 Targeted Toxic Behavior. During our thematic analysis, we defined a theme of works that focus on addressing targeted attacks, including *harassment* and *identity-based attacks*. We distinguish this behavior from common toxicity by the frequency and aim of the attack.

The first sub-theme within targeted toxic behavior is *harassment*. A clear example of this is cyberbullying of other players in online games, which is behavior that was commonly addressed in what we consider toxicity research now. The term cyberbullying as defined by Smith et al. [88] means "An aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself". In this sub-theme, we also find works such as those by Murnion et al. [64] and Vo et al. [102], who describe prediction systems that aim to detect cyberbullying through text analysis. Other work by Canossa et al. [13] analyses in-game behavior to predict multiple toxic behaviors, including harassment. Lastly, work by Märtens et al. [65] presents a prediction method to detect profane language by a player used to harass a different player in their own team.

The second sub-theme highlighted by our analysis contains works that try to address *identity-based attacks*. We differentiate intervention systems in this theme by the fact that they try to address toxic behavior aimed at a person's identity. Examples of this include sexual harassment, racism, personal insults, and discrimination. This includes work by Fox and Tang [26], who discuss the use of common in-game features like avatar selection and username choice to appear as a male player, attempting to prevent gender-based harassment. Similarly, work by Balci and Salah [5] aims to aid moderators in processing user complaints including sexual harassment. Lastly, Busch et al. [11] discuss the use of Terms of Service (TOS), End User Licence Agreements (EULA), and community guidelines to explicitly express rules to regulate behaviors including misogyny, racism, and hate speech.

5.1.2 General Toxic Behavior. The other type of intervention systems addresses *general toxic behavior*, which are acts not targeted at a specific person. Those are sometimes accepted and normalized within gaming communities. While sometimes normalized, these behaviors are often still disruptive. Clear examples of this can be undirected negative expressions to the player's own team, which lowers morale [97], or griefing, which intentionally disrupts another player's experience.

Within this theme, we defined three sub-themes. The first sub-type is *toxic communication* , such as addressed with interventions by Stepanova et al. [90], Daily [18] and Woo et al. [105], who all propose systems that detect vulgar, profane, or negative language. In the second sub-theme, we grouped toxic behaviors that are not communication , e.g., griefing, malicious behavior, and trolling, such as works by Kou and Gui [51], Prather et al. [75] and Tomkinson and van den Ende [96]. Finally, there were intervention systems that addressed general and unspecific types of toxic behaviors. Multiple interventions that belong in this sub-theme are proposed by Reid et al. [76], which include intervention systems that help players affected by any type of toxic behavior. Frommel et al. [29] proposed a prediction approach allowing subjective assessment of social interaction quality, which would work for any type of toxicity.

5.2 RQ 2: What Sites Are Researchers Using to Combat Toxicity?

Another important element of intervention systems is the entities through which they act, which we describe as the site where they act. These sites represent real-world entities and their characteristics, which can be leveraged as vectors to combat toxicity. We defined two overarching sites: *players* and *platforms*. Within this distinction, 18 intervention systems act through the *player* and their characteristics that we summarize in three sub-themes: *exposure*, *awareness*, and *mood*. Further, we

identified 24 interventions that act through the platform site. Within this theme, we defined a more complex structure, including *human moderators* and *toxic content*.

5.2.1 Exposure. In the *exposure* sub-theme, we find intervention systems that reduce exposure to toxicity for the player. Examples include multiple interventions proposed by Reid et al. [76], for example, a blocklist—enabling players to control who can interact with them—and a system that transforms toxic text into funny messages—effectively removing toxic messages before they are displayed to the player. Fox and Tang [26] discuss how female players use common in-game functionalities like avatar and username selection to mask their gender, reducing exposure to gender-based harassment.

5.2.2 Awareness. The second sub-theme we defined was *awareness*, which can be a powerful site for an intervention system, This includes approaches by Busch et al. [11], who describe terms of use and end user license agreements to regulate and govern online spaces. While these systems are defined and distributed on the platform site, their effect acts through the players' awareness of these rules and social norms in the game. Another approach is demonstrated by Kaiser and Feng [45], who show that players can be made aware of the behavior of other players through a public metric in order to inform their decision on whether to play with them.

5.2.3 Mood. Lastly, we defined *mood* as a site for intervention. Sadly, it is not always possible to prevent all toxic behavior, e.g., because of the subjective nature of what is considered toxic [6]. Therefore, there is value in creating intervention systems that help repair the mood of the affected players. In our dataset, we found approaches that directly aim to provide mood repair, such as those by Reid et al. [76] who propose presenting the player with the use of positive voice lines by in-game characters, or by displaying images of cute animals. Other approaches discussed by Reid et al. [76], Kou and Gui [52], and Kou [50] explore reporting and removing negative players from gameplay. While generally focused on sanctioning the toxic players, such reporting methods have also been shown to provide feelings of control and mood repair to those who do the reporting [76].

5.2.4 Human Moderators. On the platform site we defined a theme for systems that aid *human moderators*, who we consider part of the platform management. For example, this includes a system by Canossa et al. [13] using in-game behavior to predict toxic behavior and support the work of community managers. Furthermore, we added work by Balci and Salah [5] who aid human moderators in the time-consuming task of processing information manually.

5.2.5 Toxic Content. In the second theme on the platform site, we defined a theme for systems that act through *toxic content*. Acting through toxic content proved to be a popular research area with various approaches that we further split with another level of groupings in *chat*, *voice*, and *non-communication* channels as well as toxic *players* in general. The first group is interventions that act on the *chat* channel, e.g., the detection of toxic messages [64], detection of sexist messages [20], or detection of profanity in Korean graphemes [105]. The second grouping relates to systems that act on *voice* channels, predicting toxicity from audio features [77, 109]. In the group of *non-communication* channels, we included systems that are capable of dealing with toxicity that uses non-communication features such as in-game behavior or performance [29]. Another example is a system by Andrigueto and Araujo [4], who detect aggressiveness by analyzing feelings of fear, anger, happiness and sadness. Lastly, we grouped systems that focus directly on the toxic player. Examples include the detection of toxic usernames [18] and the removal of toxic players with a decentralized system for banning cheaters and griefers [75].

5.3 RQ 3: What Methods Are Used by Intervention Systems?

In order to guide future research into intervention systems, we deemed it important to explore how current systems try to achieve their goals. To do this, we performed another thematic analysis to find common themes in system designs. In our analysis, we defined five different themes of methods, all describing unique approaches to dealing with toxic behavior: 1) *empowering*, 2) *supporting*, 3) *priming*, 4) *sanctioning*, and 5) *detecting*.

5.3.1 Empowering. Five intervention systems in our data set use empowerment of the player, i.e., giving them tools so that they can combat the effects of toxicity themselves. The systems in this group enable the player to act against a toxic player or to reduce the chance of being targeted by toxicity. Systems that allow players to act against toxic players are demonstrated by Reid et al. [76] and Kou and Gui [52], who demonstrate systems that allow players to block misbehaving players or report their behavior to the games' moderation team. The work by Fox and Tang [26] explains in-game systems that enable players to hide their real-life identity, preventing gender-based toxicity. Lastly, the system presented by Kaiser and Feng [45] proposes an extension for the game World of Warcraft [21] that enables players to rate the behavior of other players, allowing others to see a player's social score before choosing to engage with them.

5.3.2 Supporting. Five intervention systems have supporting as their method for reducing the effects of toxicity. This method focuses on providing support to the player exposed to toxicity. This is currently only attempted by Reid at al. [76] who proposed all intervention systems in this theme. The systems include the use of positive voice lines by the in-game characters that can be requested at any time, showing pictures of cute animals to relieve stress, requesting support from one's teammates, and transforming toxic text into funny messages. The work demonstrated that all of these systems are effective in reducing the stress experienced by victims of toxic behaviors, but it also highlighted that players are not always open to being supported in such direct ways.

5.3.3 Priming. Priming is another effective method to reduce toxicity. We identified four intervention systems that rely on this method. Creating awareness of what is allowed or acceptable and what is not enables players to self-moderate and moderate each other. This is demonstrated by work by Busch et al. [11] who explore how game publishers regulate and govern their games by having players accept terms and conditions in documents like the terms of use and end user license agreements. These documents often contain the rules of a platform and state that failure to comply with these rules may end in account termination. Priming users with this knowledge aims to lead to better behavior. Another example is a system by Brewer et al. [10]. Their paper discusses the development and deployment of the GLHF (Good Luck Have Fun[2]) awareness campaign on Twitch. Through this campaign users of the platform were able to pledge to a social contract stating rules for better social interaction quality in and around video games. Those who made this pledge received a badge next to their name in the Twitch chat. This proved to be so effective that users were found to correct the behavior of other users who misbehaved but had the badge, even going so far as to report them to the GLHF team. When players were confronted with their behavior, they often showed remorse and improved their behavior afterwards, highlighting the strong effect of these social rules.

5.3.4 Sanctioning. Sanctioning is perhaps one of the first things that comes to mind when thinking about intervention systems in online games. Many online games have a way of controlling the players through sanctions. Those who misbehave often receive sanctions ranging from being disallowed to communicate to temporary removal and in some cases permanent removal from the

[2] https://www.anykey.org/pledge

game. We identified six intervention systems that rely on this method. The concept of flagging players, removing players, and moderation has been well documented by Kou and Gui [52] and Kou [50], Kou and Gui [51]. Prather et al. [75] propose a system for decentralized detection and player removal system. This is useful for multiplayer games that do not rely on a central server in their game design. Lastly, we included the system proposed by Daily [18]. This system filters chat messages and vulgar usernames. While such filtering might not immediately seem like a sanction, it does take away privileges of communication or freely choosing names.

5.3.5 Detecting. The last method we defined is detection. Detection is one of the key steps for any intervention system. It is also the largest theme we defined, with 18 intervention systems relying on this method. For any subsequent intervention to occur, there first has to be a form of detection. This is commonly achieved through the use of machine learning algorithms, however, detection also includes human input through a player or moderator. Compared to the other categories in this thematic analysis, the systems in this grouping only perform this detection step and therefore do not help combat toxicity without further steps. While they do not contain an action element, their contributions are valuable. For example, Canossa et al. [13] demonstrate a system that can predict toxic behavior based on gameplay actions. They go as far as to state that proactive action would be possible, eliminating the need for a validation step by human moderators. Frommel et al. [29] hint at similar functionality through continuous monitoring of social interaction quality and Reid et al. [77] also mention the possibility of proactive action. While all of these works state that this concept is untested, it does highlight that there is research interest. While proactively intervening in potentially toxic behavior has potential downsides (e.g., removing falsely flagged and innocuous content or preemptively scrutinizing non-toxic players), predictors accurate enough to perform such a task would be extremely useful when implemented in a way that they can prevent harm.

5.4 RQ 4: Are Interventions Applied Before or After Harm From Toxic Behavior Occurs?

To assess when intervention systems are applied, we created a deductive coding scheme during the screening phase. We coded systems with *before* or *after*, depending on when the system would act. We found that 31 intervention systems act *after* toxicity occurs. The remaining five perform this task *before*. We only coded five intervention systems as acting *before* toxicity, e.g., including work by Busch et al. [11] who describe legal agreements that aim to nudge players towards better behavior before they are toxic. Similarly, Fox and Tang [26] describe the act of gender masking as a coping strategy that women use against harassment. We included this as a system that acts *before* toxicity because it can help avoid exposure. It does make for an exceptional case because gender masking is performed before toxic behavior occurs, however, it is likely a reactive action to prior experienced toxic behavior. We observed that the majority of intervention systems take action *after* toxicity, which is easier because there is a clear trigger for intervention. These interventions include artificial intelligence (AI) systems that detect if toxicity has occurred (e.g., [58, 64]) and the systems that provide mood repair after exposure to toxicity [76]. Furthermore, we coded three approaches for toxicity detection [13, 29, 77] as *after* toxicity, although the papers explain that their systems could be used in a proactive manner (i.e., predicting toxicity before it happens). As described in Section 5.3.5, these papers state that this concept is possible, but none of the papers demonstrate it.

5.5 RQ 5: How Are Researchers Evaluating the Effectiveness of Their Intervention Systems?

In order to answer this research question, we use three sub-questions: *"What metric do authors use to evaluate their interventions?"*, *"Do Authors Evaluate Existing or Novel Interventions?"*, and *"Do Authors Evaluate Their Interventions in Realistic Settings?"*.

5.6 RQ 5.1: What Metric Do Authors Use to Evaluate Their Interventions?

A critical component of system design is its evaluation. This step allows the creators of a system to validate whether design criteria and overall quality goals are met, and to evaluate and quantify the effectiveness of the implemented solution. Intervention systems are no exception to this. An effective intervention could have the potential to improve the experience of video game players worldwide. Hence, valid and insightful evaluations are greatly desirable. We defined different evaluation approaches: *model performance*, *user studies*, *simulations*, and *content and system analysis*. In addition, some systems were *not evaluated*.

5.6.1 Model Performance. We found that 17 intervention systems are evaluated based on model performance. As noted, many interventions in our data set are detection approaches using AI/machine learning models on chat data [8, 98]. Such approaches are usually evaluated with model performance metrics, such as accuracy, precision, or F-measure, a single score depicting a balanced calculation of precision and recall [101]—metrics that can be used to quantify the performance of a classification model. Another metric used is the diagnostic odds ratio, which is another assessment of effectiveness for binary classification problems, and is found in an intervention system performing semantic analysis of in-game chat for cyberbullying, presented by Murion et al. [64].

5.6.2 User Studies. 17 intervention systems in our data set were evaluated with a user study. 10 evaluations use qualitative methods with different methodological approaches. Kou and Gui [51] performed an analysis of forum threads about moderation on the official League of Legends [31] forum, exploring the concept of governing the online game with humans and with AI. Another study by Reid et al. [76] let players describe what they desired from in-game support tools. Lastly, Kou [50] performed a qualitative analysis of player discourse on the concept of permanent bans in League of Legends. 7 intervention systems were evaluated with a quantitative user study. For example, Reid et al. [76] the effectiveness of their interventions quantitatively measuring mood before and after using an intervention focused on mood repair. Fox and Tang [26] assessed women's experience of harassment and the use of coping strategies in an online game context with a survey and Brewer et al. [10] report findings from a community-driven intervention deployed in the wild amongst 370.000 gamers and evaluated with usage statistics. There are some cases where user studies also included mixed methods (e.g., [76]), which is valuable because it can provide deeper insights. For example, such an evaluation showed that an intervention is effective but potentially still rejected by some players as silly [76].

5.6.3 Simulations. Two intervention systems were evaluated through simulations [45, 75]. As opposed to using real players for their evaluation, these system designs allowed for computational testing of their functionality. In both cases, a simulated environment was used to validate the performance and whether design criteria were met.

5.6.4 Content and System Analysis . Two of the intervention systems were evaluated through a content and system analysis. Busch et al. [11] analyze different legal documents and the implementation of their policies. Their assessment analyses how two different companies justify

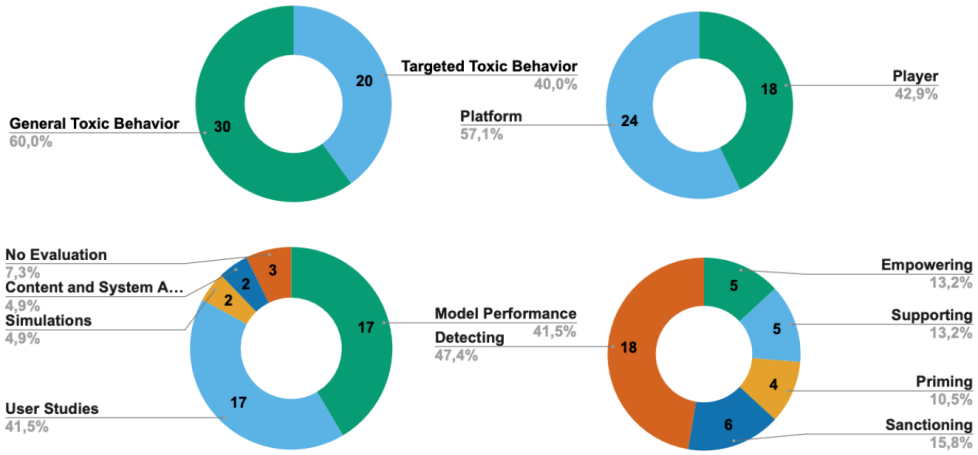

Fig. 2. Distribution of intervention systems across RQs 1–3 and RQ 5.1.

and legitimize their legal efforts. Tomkinson and van den Ende [96] have analyzed the effect of intervention through positive reward in the endorsement system of the game Overwatch.

5.6.5 No Evaluation. For 3 interventions, the corresponding paper did not include an evaluation. Some interventions were presented in non-archival publications or as preliminary findings without a full evaluation, e.g., Daily [18] and Stepanova [90].

5.7 RQ 5.2: Do Authors Evaluate Existing or Novel Interventions?

In the corpus, 28 interventions were a *novel* way of addressing toxic behavior, meaning that they proposed a new intervention system. The remaining 8 studies represented an *existing* intervention system, meaning that they evaluated a system that is actively being applied in a gaming context. We observed that the majority of the collected papers proposed novel interventions (e.g., automatically detecting cyberbullying comments [102] or using conversational networks for online moderation [72]) while existing systems (e.g., the endorsement system in Overwatch [96]) are less commonly studied.

5.8 RQ 5.3: Do Authors Evaluate Their Interventions in Realistic Settings?

In order to assess how realistic the evaluation processes were for intervention systems, we coded systems based on two factors. We define a realistic evaluation as one that most closely resembles a real-world situation, in which an intervention system would act. For each intervention system, we coded whether the intervention system was based on a commercial setting and whether the intervention system had been evaluated with players. We observe that 22 intervention systems are *based on commercial settings*, either using data that originates from commercial settings or assessing systems used within commercial settings (e.g., [20, 45]). The remaining 14 intervention evaluations were *not based on commercial settings* (e.g., [75, 90]). Out of the 36 intervention systems, 12 are *evaluated with players* (e.g., [51, 76]) while the majority of the intervention systems ($n = 24$) was *not evaluated with players* (e.g., [58, 87]). By combining these two factors, we observed that only five of the interventions are *evaluated with players* and *based on commercial settings* ([10, 26, 50–52]). A full overview of our findings is reported in Table 6. The relationship between *evaluation with players* and *based on commercial settings* was not significant, $\chi^2(1, N = 36) = 2.864$, $p = .091$.

5.9 Identifying Patterns and Research Gaps Through Cross-Tabulation

To provide concrete guidance for future work into novel intervention systems we cross-tabulated data from the different themes, identifying areas of interest in the current design space. Table 7 to Table 9 allow for a categorical comparison between three of the four key themes of intervention systems discussed in this paper: types of toxicity, toxicity sites, and intervention methods. We did not include evaluation approaches here, as these do not define a unique aspect of an intervention system's design, but included them in the Appendix for interested readers.

In Table 7, we observe that empowerment and support are underrepresented intervention methods for all types of toxicity, except for unspecified toxicity. Further, we see that detection of harassment, identity-based attacks, and toxic communication are well studied areas in the design space. Table 8 shows the cross-tabulation between intervention methods and intervention sites. We observe that none of the intervention systems in our data have applied supporting, priming, or empowering through the site of the platform. For example, no work has looked into providing support, e.g., mood repair, to human moderators who are exposed to a lot of toxicity by the nature of their jobs. Therefore, they may also benefit from support mechanisms. Lastly, we see that priming of a player's mood has not been attempted. Priming players with positivity might be able to reduce toxic outbursts. When comparing types of toxicity and intervention sites (see Table 9), we observe that *targeted attacks (harassment* and *identity-based attacks)* are only well studied in chat content. There is not much research about such targeted attacks in the voice communication and non-communication content sites, indicating a gap. This is problematic because targeted attacks can also occur in such channels. In its latest report, the Anti-Defamation League even states that harassment amongst young people (aged 10-17) most commonly occurs through voice channels, with a rise from 39% up to 53% between 2021 and 2023 [57]. On a positive note, we see strong research interest in communication based toxicity in chat channels. Further, we can see interest in player-awareness-based intervention systems for all types of toxicity (see Table 9).

6 Discussion

6.1 Summary of Findings

In summary, intervention systems in the literature aim to combat two main types of toxicity: *targeted toxic behavior* and *general toxic behavior*, act through *players* and *platforms* to achieve their goals, and employ five methods that intervention systems use to achieve their goal: *empowering, supporting, priming, sanctioning*, and *detecting*. We categorized intervention systems based on when they act and found that most current systems act only *after* harm has occurred. Finally, we explored if and how these intervention systems were evaluated. We found that most systems are *novel* solutions while there is a gap regarding the evaluation of *existing* interventions (e.g., bans, reporting) as well as evaluations in the context of commercial settings and with players. We synthesized these findings across our analysis dimensions and provided an overview that can be used as a design space of intervention systems.

6.2 How Current Intervention Systems Work

We categorized systems based on where they act (sites, RQ2) and how they work (methods, RQ3). Our analysis provides an overview of the state of the literature and gaps within it.

6.2.1 Sites. Regarding sites, we categorized interventions in two themes: *players* and *platforms*, which aligns with previous findings in content moderation research [82] As toxic behavior mostly originates from players and is aimed at players, one might expect that the player theme would be the largest. In practice, only 43 percent of intervention systems act through the players, including both the toxic players (e.g., informing players about their own negative behavior [90]) and the targeted

players (e.g., through mood repair [76]). We believe that there is potential in systems that act through players. For example, the mood repair systems proposed by Reid et al. [76] were effective at providing positive effects. Furthermore, creating player awareness has been demonstrated to have positive effects on behavior [10]. Lastly, enabling players to control their own exposure to toxic behavior can facilitate personal boundaries for these behaviors. This allows players to self-regulate their exposure to toxicity, reducing the risk of community pushback when strict interventions are imposed on the entire community of a game by its developer [80]. Our findings demonstrate that there could be more approaches that focus on the player. This could empower them to control their own gaming experience, stimulate better behavior and provide support when all other methods fail.

On the other hand, the second theme comprised of interventions that act through the platform site. Within that, a significant majority of the interventions fit in a sub-theme that acts through toxic content. Most of these systems act on the toxic content with a focus on text-based communication and only three systems acting on voice-based toxic content. This is interesting, as voice communication is important and widely used in high-paced and high-stakes gaming situations. Specifically, these situations are known to foster toxic behavior more than calmer gameplay [1], highlighting the need for interventions applied to voice channels (and non-communication).

6.2.2 Methods. Considering the methods of interventions, nearly half of the systems are approaches that focus on *detecting* toxicity. Importantly, the overwhelming majority of detection systems focus on chat-based data (see Table 8), which is certainly valuable, but limited in scope. This trend can potentially be explained by advances in text analysis and the prevalence of text as a communication channel in many games. However, toxicity also happens in other channels like voice communication or through in-game behaviors, for which there are far fewer intervention methods. Such approaches would be valuable to pursue to bridge the gap, while recent advances in novel AI approaches (e.g., multimodal foundation models like Gemini [32]) may be viable to help with detection in video or audio. While toxicity detection is essential, it does not help combat toxicity alone without subsequent action like filtering. One such follow-up is *sanctioning*, which is another theme for methods used in the dataset. Together, detecting and sanctioning toxicity can be an effective strategy for combating toxic behaviors. Sanctions comprise restrictions limited to a communication channel (e.g., text bans), time (e.g., temporal bans), or play mode (e.g., low-priority matchmaking queues) as well as permanent removal from games. While such punishments of course are essential for removing players who are continuously toxic, it is unclear whether (temporary) sanctions are effective at changing a toxic player's future behavior. We do believe that normalization of such behavior could be fought using more enforcement and the stimulation of more positive social interaction in games. Another interesting theme featured approaches focused on *priming*, which includes legal contracts [11] and awareness campaigns [10]. This method can target all types of players and potentially effectively and proactively nudge players toward better behavior. Yet, such methods are commonly difficult to implement in an effective way: For example, agreements like codes of conduct are often inaccessible to users [33], while awareness campaigns require reaching lots of players, a feat out of reach for many research labs. Further, it is also non-trivial to measure the effectiveness of such solutions in a real-world setting. The two remaining approaches center on the targeted players. Several approaches focus on *empowering* players in that it gives them additional tools so that they can combat toxicity themselves. For example, we include user-centered approaches to moderation, such as blocking, which allows users to take control over their own exposure. Such control-over-exposure approaches have recently been suggested as a way to combat toxicity in games [27] and studied as a way of personalized moderation in social media [42, 43]. Finally, there are some approaches for *supporting* targeted players, such as via mood repair [76]. From these findings we can argue that the application of different intervention methods can help

combat different types of toxic behavior in online games, which is in line with findings in content moderation research [83].

6.3 Future Toxicity Interventions

Our cross-tabulation of research results allows us to make suggestions for future intervention systems by highlighting potential research gaps in the current design space of intervention systems. For example, we believe that both victims of toxic communication and targeted attacks might benefit from support and empowerment, for whom it could possibly reduce direct and long-lasting impact. This area is currently underpopulated in the design space. Further, we observed that while human moderators are perhaps most commonly exposed to toxicity by the nature of their jobs, there is no approach to provide support for them helping them deal with such content. This gap in the design space, combined with existing literature from the social media space on moderators' mental wellbeing [79] highlights the need for such support mechanisms. Overall, we found that the various methods used in the reviewed papers demonstrate a broad and diverse research scope.

6.4 Different Interventions For Different Types of Toxicity

In our analysis, we differentiated intervention research focused on specific types of toxicity, including *general toxic behavior* and *targeted toxic behavior*. Considering the nuances between different types of toxicity is essential for implementing interventions. For example, general toxicity is usually undesirable in a games context but also normalized and sometimes acceptable in certain contexts. For example, in many (adult) contexts, using a swear word such as "F***k yes" can be considered acceptable. Recent work has argued for the potential of individual control over exposure as a way to combat toxicity [27]. Such approaches would be considered interventions that decrease *exposure*, while also providing nuance to interventions, e.g., allowing for trash talk in games for those that do not consider them toxic [29]. This is in line with other recent research that found that a player's offline culture has a significant impact on the occurrence and perception of toxic behavior [48], increasing the difficulty of assessing what is acceptable and what is not. On the other hand, *targeted toxic behavior* like *harassment* and *identity-based attacks* are never acceptable. This should be considered in intervention design, e.g., with detection approaches or sanctions that can differentiate between them. This is also important because some approaches for *general toxic behavior* may not work when applied to *targeted toxic behavior*, e.g., a word filter that a harasser can easily circumvent. On the other hand, such *targeted toxic behavior* may be suitably addressed by other approaches like *awareness* campaigns, similar to training interventions for (non-game) cyberbullying [16, 95].

In the same way, we also think that it is necessary to consider differences in where toxicity happens from an interventions perspective. For example, approaches that identify toxicity in communication, e.g., [87, 109], may not be applicable to detect non-communication toxicity such as griefing, which can happen through gameplay instead of chat. Due to the wide variety of communication options in games and the wide prevalence of toxicity among those channels, more research must examine how to detect and combat toxicity in different channels and sites. Systems with unspecified toxicity aims can serve multiple roles, for example enabling players to manage their experience around toxic players [45], or interventions such as removing a player from the game based on various types of undesirable behavior [50].

6.5 There *Is* Glory in Prevention

In our analysis of how intervention systems act, we assessed when they take action. Most of the intervention systems in the literature act only *after* harm has already occurred. This is natural, as most systems intervene as a reaction to toxic behavior, e.g., detecting, sanctions, or aftercare.

Intervention systems that intervene before toxicity have huge potential because they could prevent harm instead of mitigating it. We recognize that this is challenging because it relies on approaches that reliably predict toxicity to trigger intervention (which is hindered by the lack of tools for this purpose and the subjectivity of toxicity [28]), or the development of systems that lead to more positive communities and player interactions [96] (which is difficult due to the increasing normalization of toxicity [6]). While certainly difficult to implement and in need of more research, proactive and preventative approaches that act before toxicity are promising, ideally preventing harm before it occurs. This could involve approaches to reduce exposure [27] or approaches that actively manage the players' mood during gameplay. For example, an intervention system could predict frustration using low-level interaction trace data from input devices (e.g., [22, 30, 61]), behavioral metrics like body posture (e.g., [78]), or game performance measures (e.g., [29]) and intervene before it progresses to raging. This intervention could deliver distractions to the toxic player in-game (e.g., provide automated suggestions for changing match strategies), external devices (e.g., interactive stress relief devices similar to stress balls), or suggestions for emotional control (e.g., controlled breathing).

6.6 Developing and Evaluating Interventions

As the next step of our literature study, we assessed the evaluation practices of the proposed intervention systems. The metrics used to evaluate intervention systems align with five themes, with a majority of interventions fitting into two of them. *User studies* and *model performance* both make up 42% of the evaluations. It is promising that many intervention systems are evaluated with user studies. This is valuable as it can provide a measure of external validity, as this type of evaluation incorporates the players' experiences, and should therefore be representative of the population that we want to protect with interventions. Another large part of evaluation approaches is based on *model performance*. Such evaluations are applied to approaches that use models to detect toxicity, for which they are essential for providing evidence for the validity of a prediction model. However, there is a gap in evaluations connecting such detection approaches together with other moderation approaches like sanctioning or filtering and how they affect players.

In our analysis, we found that most intervention systems are *novel*, which is encouraging. There is a lot of value in the creation of new systems. Such research can help improve the current state of the art and accordingly is also the core of artifact contributions in HCI [104]. However, there is also value in assessments of existing systems to better understand existing strategies and to evaluate their effectiveness. Such work can provide valuable insights, such as highlighting that reporting is often misused [52], while also arguing the potential benefits like mood repair through the act of reporting [76]. Especially considering external validity, this points to a gap in the literature about the evaluation of already existing interventions. Ultimately, the game industry also works hard on implementing systems that combat toxicity. However, we do not know much about the design and evaluation of such solutions. Future work may benefit from contributions that investigate toxicity interventions that are commonly used in games (e.g., more research on reporting, muting, or blocking) or larger systems (e.g., Dota 2's new player behavior system[3]). Regarding the question if authors evaluate their intervention systems in a *realistic setting*, we think there is a potential to improve the external validity of intervention systems. In our analysis, we found that only five interventions were evaluated with players and based on commercial settings. Such evaluations are beneficial because they consider the dynamics of play. Evaluation of interventions outside of realistic contexts may not fully capture how they would work in existing commercial games

[3]https://www.dota2.com/summer2023: personalized matchmaking; new reporting system; recognizing good reports; real-time processing of toxic chat; and behaviour and communication scores

with other expectations and norms. Similarly, involving players in the evaluation is essential for assessing effectiveness and user acceptance. For example, some study participants considered approaches to mood repair as silly, highlighting the need for appropriate integration and subsequent evaluation [76]. Thus, evaluations with players in *commercial settings* are particularly insightful and valuable, such as those focused on systems in League of Legends [50–52] and coping strategies that could inform interventions [26]. Only Brewer et al. [10] have proposed a novel system that was evaluated with players and in commercial settings, namely a public awareness campaign working through Twitch and not in gameplay. Out of the five intervention systems acting *before* toxicity, three include evaluations of existing systems [11, 26, 96]. The fourth intervention [18] dates back to 2006 and was part of an online network that no longer exists. Since then, only Brewer et al. [10] proposed an approach for an intervention system that could prevent harm. To summarize, evaluation in intervention research could be stronger by including players directly and applying them in commercial settings. Such evaluations are essential for estimating the effect in a real-world context. While understanding the complexity of such an evaluation, we strongly suggest such approaches for a comprehensive evaluation of future intervention research while this would be valuable for existing interventions too.

6.7 Closing Gaps Between Academia and Industry

Our analysis showed many novel intervention systems, but we rarely see them being tested in commercial games. We believe that there is mutual gain to be had in collaborating on projects. For academic researchers, collaboration can enable better research through realistic and current data, evaluation studies with higher external validity, and societal impact through embedding intervention approaches into contexts where they can combat harm. On the other hand, the game industry can benefit from academic research that provides rigorous empirical evidence about the effectiveness of interventions, algorithms and system components, and design considerations for intervention tools. For researchers in both academia and industry, we provide an overview of currently existing intervention systems, their goals, methods and evaluation approaches. This overview can guide research by highlighting the design space of interventions from prior research. Further, we provide insights into the areas of the design space that are currently unexplored, guiding future work. Ultimately, everyone interested in reducing the harm of toxicity will benefit from more collaboration and applied interventions integrated into commercial games, including researchers, game makers, community moderators, and ultimately players.

7 Limitations & Future Work

There are a few limitations to our study. First, as a systematic literature review, our literature search is limited to the scope of our goals and the search terms that we used. For example, work by Grace et al. [33] that describes codes of conduct is not in our data set, because no toxicity keyword is mentioned in the title or abstract. Similarly, work by Sengün et al. [84, 85] that may be used by an intervention was out of the scope of our search strategy. We excluded work like Komaç et al. [46] that presents a standalone serious game to increase awareness about trolling. While the approach of a serious game can help combat toxicity, our objective was to explore interventions that are embedded in online games or game environments where toxicity occurs. A serious game designed to raise awareness of toxicity, or promote effective self-regulation, is not a component of the game in which toxicity occurs. Further, we did not include non-game interventions in our review, such as approaches applied to social media that provide feedback to toxic users [106], which may have potential for application in online games.

Second, we used active learning to assist in the abstract screening. This is a novel approach that can substantially reduce time and effort in screening irrelevant papers [62, 100]. There are early

explorations the performance of active learning methods in general [110] and ASReview specifically [100] that suggested very high accuracies [24, 99]. However, this approach still remains less well understood compared to traditional, manual screening. In comparison to other examples (e.g., [100]), we used a conservative stopping threshold (10% of data without an inclusion) to mitigate potentially missing references, while relying on the algorithm to support the assessment of records that have a high likelihood of being included.

Third, we note that our analysis method was an interpretive process that clustered and aligned the papers based on their characteristics through multiple iterations and discussions (with the first and last authors, and then in later stages with the rest of the author team). Care was taken to keep the papers accurately presented and aligned with the broader literature, but of course, the resulting categorization (i.e., themes) is shaped by the authors' understanding of the field as a whole. To provide context on this [60, 93], we note that the first, third, and last author were closely familiar with toxicity literature and games in general, while the other authors involved in the later discussion provided games research backgrounds in other areas of the field. Our analysis was guided by shared knowledge covering a variety of different game genres and gaming spaces, computer science, information science, and HCI backgrounds and involvement in HCI and Games communities. Through personal experience, prior research of the authors, and examples in the literature, we interpreted the information to define our initial themes. Theme generation was further progressed by discussions amongst the entire team. Different researchers may generate slightly different themes, but we believe that the final results are well-suited to constitute a foundational understanding of toxicity interventions. Through this method we have illuminated the intervention system space on a broad level. Future work can extend on this by exploring individual intervention systems on a more in-depth level using existing related frameworks as deductive coding schemes, such as the trade-off-centered framework of content moderation by Jiang et al. [44] or the different emotions, perspectives, and behaviours experienced by users in content moderation as identified by Ma et al. [59].

Finally, as a literature review of research papers, we cannot provide insights about interventions applied in the games industry. Several companies already apply interventions in their game environments (e.g., in Overwatch [39] or Dota 2, or on FACEIT [14, 15]). In this paper, we do not include this industry perspective on interventions, but highlight it as an interesting direction for future work, e.g., analyzing blog posts or transparency reports discussing interventions, release notes that describe new approaches, or community discourse on platforms in online communities. Further, it would be valuable to interview game industry representatives about their opinions on toxicity interventions.

8 Conclusion

In this paper, we presented a systematic literature review on toxicity interventions in online games, making several key contributions. First, we show that intervention systems predominantly emphasize addressing toxicity within communication channels, leaving other channels, such as gameplay-related toxicity, largely unexplored. Notably, the majority of existing systems concentrate on addressing in-game chat; few intervention systems specifically address toxicity through voice channels. These findings highlight a research gap in designing interventions for one of the most likely places in which toxic behaviors occur. Moreover, our findings demonstrate that a healthy distribution intervention systems operate within the confines of the gaming platform itself (e.g., platform-side detection of toxicity) and interventions that act via the player. For example, this includes empowering players to combat the effects of toxicity themselves, creating awareness of how their communications and behaviours are received, or repairing mood after exposure has occurred. Additionally, our findings reveal that a substantial amount of intervention systems only

perform toxicity detection, while other steps (e.g., subsequent sanctions) are not explored to the same degree. We advocate for increased attention in future research to methods that combine detection with active intervention. We found that there are far fewer intervention systems that act before toxicity affects another player than interventions acting before toxicity, pointing to a gap in approaches that can prioritize harm prevention over mitigation strategies. In considering the approaches for evaluating the efficacy of the interventions, our findings demonstrate that user study results and model performance are the most common metrics employed. While user studies offer high external validity, we acknowledge that their execution can be challenging, and that approaches to quantifying or qualifying harm prevention are still in their infancy. Lastly, our analysis highlights limited validation approaches that gather data from players or that are based on commercial games and game environments, pointing to an opportunity and a need to enhance the external validity of toxicity intervention research. In conclusion, our study identifies multiple research gaps in toxicity intervention research, providing valuable insights into the current state of the field. These findings offer new insights related to the designs, goals, and evaluations of toxicity interventions, and provide valuable guidance to researchers interested in developing and assessing approaches to combat toxicity in multiplayer online gaming.

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*. ACM, Melbourne VIC Australia, 365–372. https://doi.org/10.1145/3270316.3271545

[2] Icek Ajzen. 1985. From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior*. Springer, 11–39.

[3] R. Algozzine and P. Kay. 2002. *Preventing Problem Behaviors: A Handbook of Successful Prevention Strategies*. SAGE Publications. https://books.google.nl/books?id=-kQy3rJ1vz8C

[4] G.R. Andrigueto and E. Araujo. 2020. Fuzzy aggressive behavior assessment of toxic players in multiplayer online battle games. In *IEEE International Conference on Fuzzy Systems*, Vol. 2020-July. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/FUZZ48607.2020.9177560

[5] Koray Balci and Albert Ali Salah. 2015. Automatic Analysis and Identification of Verbal Aggression and Abusive Behaviors for Online Social Games. *Comput. Hum. Behav.* 53, C (Dec. 2015), 517–526. https://doi.org/10.1016/j.chb.2014.10.025

[6] N.A. Beres, J. Frommel, E. Reid, R.L. Mandryk, and M. Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445157

[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 (Sept. 2023), 558–587. https://doi.org/10.1145/3611041

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

[9] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[10] Johanna Brewer, Morgan Romine, and T. L. Taylor. 2020. Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, New York, NY, USA, 757–769. https://doi.org/10.1145/3357236.3395514

[11] T. Busch, K. Boudreau, and M. Consalvo. 2015. *Toxic gamer culture, corporate regulation, and standards of behavior among players of online games*. Taylor and Francis. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027273907&doi=10.4324%2f9781315748825-13&partnerID=40&md5=a597d7ef15159efc423886e9f3808c7b

[12] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 56, 3 (June 2022), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

[13] A. Canossa, D. Salimov, A. Azadvar, C. Harteveld, and G. Yannakakis. 2021. For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay. *Proceedings of the ACM on Human-Computer Interaction* 5, CHIPLAY (2021). https://doi.org/10.1145/3474680

[14] Joel Chapman and FACEIT. 2020. Minerva Gets Ears: Tackling Toxicity in Voice Chat. https://blog.faceit.com/minerva-gets-ears-tackling-toxicity-in-voice-chat-3eb80471d005
[15] Joel Chapman and FACEIT. 2021. Fighting Abusive Behaviour — Product Update. https://medium.com/faceit-blog/fighting-abusive-behaviour-product-update-d39e490c00ce
[16] Enrique Chaux, Ana María Velásquez, Anja Schultze-Krumbholz, and Herbert Scheithauer. 2016. Effects of the cyberbullying prevention program media heroes (Medienhelden) on traditional bullying. *Aggressive behavior* 42, 2 (2016), 157–165.
[17] J.A. Cornel, C. Christian Pablo, J.A. Marzan, V. Julius Mercado, B. Fabito, R. Rodriguez, M. Octaviano, N. Oco, and A.D. La Cruz. 2019. Cyberbullying Detection for Online Games Chat Logs using Deep Learning. In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2019*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/HNICEM48295.2019.9072811
[18] G. Daily. 2006. A case of delivering family-friendly entertainment. *EContent* 29, 4 (2006), 45–47. https://www.scopus.com/inward/record.uri?eid=2-s2.0-33750903305&partnerID=40&md5=4e3c40c38b035dd34f6ab782e9aeaa80
[19] Julian Dibbell. 1994. A Rape in Cyberspace; or, How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database into a Society. In *Flame Wars*, Mark Dery (Ed.). Duke University Press, 237–261. https://doi.org/10.1215/9780822396765-012
[20] A. Ekiciler, İ. Ahioğlu, N. Yıldırım, İ.İ. Ajas, and T. Kaya. 2022. The Bullying Game: Sexism Based Toxic Language Analysis on Online Games Chat Logs by Text Mining. *Journal of International Women's Studies* 24, 3 (2022). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134998324&partnerID=40&md5=6e6bc81ee9201371bc0a80eb31071f8d
[21] Blizzard Entertainment. 2004. *World of Warcraft*. Game [PC].
[22] Clayton Epp, Michael Lippold, and Regan L Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*. 715–724.
[23] Ted Feinberg and Nicole Robey. 2009. Cyberbullying: Intervention and prevention strategies. *National Association of School Psychologists* 38, 4 (2009), 22–24.
[24] Gerbrich Ferdinands, Raoul Schram, Jonathan De Bruin, Ayoub Bagheri, Daniel L. Oberski, Lars Tummers, Jelle Jasper Teijema, and Rens Van De Schoot. 2023. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. *Systematic Reviews* 12, 1 (June 2023), 100. https://doi.org/10.1186/s13643-023-02257-7
[25] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting online harassment of women journalists. *Mediating misogyny: Gender, technology, and harassment* (2018), 311–332.
[26] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *NEW MEDIA & SOCIETY* 19, 8 (Aug. 2017), 1290–1307. https://doi.org/10.1177/1461444816635778
[27] Julian Frommel and Regan L. Mandryk. 2023. Individual Control over Exposure to Combat Toxicity in Games. *ACM Games* 1, 4, Article 24 (dec 2023), 3 pages. https://doi.org/10.1145/3633768
[28] 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* (Bremen, Germany) *(CHI PLAY '22)*. Association for Computing Machinery, New York, NY, USA, 263–265. https://doi.org/10.1145/3505270.3558359
[29] Julian Frommel, Valentin Sagl, Ansgar E. Depping, Colby Johanson, Matthew K. Miller, and Regan L. Mandryk. 2020. Recognizing Affiliation: Using Behavioural Traces to Predict the Quality of Social Interactions in Online Games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376446
[30] Julian Frommel, Claudia Schrader, and Michael Weber. 2018. Towards emotion-based adaptive games: Emotion recognition via input and performance features. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 173–185.
[31] Riot Games. 2009. *League of Legends*. Game [PC].
[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[33] Thomas D. Grace, Ian Larson, and Katie Salen. 2022. Policies of Misconduct: A Content Analysis of Codes of Conduct for Online Multiplayer Games. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (Oct. 2022), 1–23. https://doi.org/10.1145/3549513
[34] Kate Grandprey-Shores, Yilin He, Kristina L. Swanenburg, Robert Kraut, and John Riedl. 2014. The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, Baltimore Maryland USA, 1356–1365. https://doi.org/10.1145/

2531602.2531724
[35] K. Grieman. 2019. Lakitu's world: Proactive and reactive regulation in video games. *Interactive Entertainment Law Review* 2, 2 (2019), 67–77. https://doi.org/10.4337/ielr.2019.02.02
[36] James Grimmelmann. 2017. The Virtues of Moderation. https://doi.org/10.31228/osf.io/qwxf5
[37] Shengbo Guo, Scott Sanner, Thore Graepel, and Wray Buntine. 2012. Score-Based Bayesian Skill Learning. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I (ECMLPKDD'12)*. Springer-Verlag, Berlin, Heidelberg, 106–121.
[38] Robert Hahn, Dawna Fuqua-Whitley, Holly Wethington, Jessica Lowy, Alex Crosby, Mindy Fullilove, Robert Johnson, Akiva Liberman, Eve Moscicki, LeShawndra Price, Susan Snyder, Farris Tuma, Stella Cory, Glenda Stone, Kaushik Mukhopadhaya, Sajal Chattopadhyay, and Linda Dahlberg. 2007. Effectiveness of Universal School-Based Programs to Prevent Violent and Aggressive Behavior. *American Journal of Preventive Medicine* 33, 2 (Aug. 2007), S114–S129. https://doi.org/10.1016/j.amepre.2007.04.012
[39] Iain Harris. 2020. Toxicity in Overwatch has seen an "incredible decrease" due to machine learning. https://www.pcgamesn.com/overwatch/toxic-behaviour-machine-learning.
[40] Janne Huuskonen. 2022. Toxicity and the hidden dangers of shoddy moderation. https://utopiaanalytics.com/toxicity-and-the-hidden-dangers-of-shoddy-content-moderation/
[41] I.O. Ididi, S. Hassan, A.A.A. Ghani, and N.M. Ali. 2017. Excessive and addictive gaming control using counselling agent in online game design. In *AIP Conference Proceedings*, Vol. 1891. American Institute of Physics Inc. https://doi.org/10.1063/1.5005398
[42] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (oct 2023), 33 pages. https://doi.org/10.1145/3610080
[43] Shagun Jhaver and Amy X Zhang. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* (Dec. 2023). https://doi.org/10.1177/14614448231217993
[44] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (Feb. 2023), 1–34. https://doi.org/10.1145/3534929
[45] Edward Kaiser and Wu-chang Feng. 2009. PlayerRating: A Reputation System for Multiplayer Online Games. In *Proceedings of the 8th Annual Workshop on Network and Systems Support for Games (NetGames '09)*. IEEE Press.
[46] G. Komaç and K. Çağıltay. 2021. Raising Awareness Through Games: The Influence of a Trolling Game on Perception of Toxic Behavior. *Springer Series in Design and Innovation* 13 (2021), 143–154. https://doi.org/10.1007/978-3-030-65060-5_12
[47] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *INTERNET RESEARCH* 30, 4 (Aug. 2020), 1081–1102. https://doi.org/10.1108/INTR-08-2019-0343 Place: HOWARD HOUSE, WAGON LANE, BINGLEY BD16 1WA, W YORKSHIRE, ENGLAND Publisher: EMERALD GROUP PUBLISHING LTD Type: Article.
[48] Bastian Kordyaka, Solip Park, Jeanine Krath, and Samuli Laato. 2023. Exploring the Relationship Between Offline Cultural Environments and Toxic Behavior Tendencies in Multiplayer Online Games. *ACM Transactions on Social Computing* 6, 1-2 (June 2023), 1–20. https://doi.org/10.1145/3580346
[49] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3410404.3414243
[50] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). https://doi.org/10.1145/3476075
[51] Y. Kou and X. Gui. 2017. When code governs community. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2017-January. IEEE Computer Society, 2056–2064. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108240654&partnerID=40&md5=83c3ad3957b8bdbc96aa5a84a4492893
[52] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445279
[53] Rachel Kowert. 2020. Dark Participation in Games. *Frontiers in Psychology* 11 (Nov. 2020). https://doi.org/10.3389/fpsyg.2020.598947
[54] M. Köles and Z. Péter. 2016. "Learn to play, noob!": The identification of ability profiles for different roles in an online multiplayer video game in order to improve the overal quality of the new player experience. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. 000271–000276. https://doi.org/10.1109/CogInfoCom.2016.7804560

[55] Anti-Defamation League. 2021. Hate is No Game: Harassment and Positive Social Experiences in Online Games 2021. https://www.adl.org/resources/report/hate-no-game-harassment-and-positive-social-experiences-online-games-2021
[56] Anti-Defamation League. 2022. Hate Is No Game: Hate and Harassment in Online Games 2022 | ADL. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022
[57] Anti-Defamation League. 2024. Hate Is No Game: Hate and Harassment in Online Games 2023 | ADL. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023
[58] Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *DECISION SUPPORT SYSTEMS* 113 (Sept. 2018), 22–31. https://doi.org/10.1016/j.dss.2018.06.009
[59] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 278 (oct 2023), 30 pages. https://doi.org/10.1145/3610069
[60] Tim May and Beth Perry. 2014. *Reflexivity and the practice of qualitative research.* Vol. 109. Sage, Los Angeles, USA, Chapter 8, 109–122.
[61] Matthew K Miller and Regan L Mandryk. 2016. Differentiating in-game frustration from at-game frustration using touch pressure. In *Proceedings of the 2016 ACM international conference on interactive surfaces and spaces*. 225–234.
[62] Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 51 (Oct. 2014), 242–253. https://doi.org/10.1016/j.jbi.2014.06.005
[63] Emily Morrow. 2022. All Anti-Toxicity Changes from Overwatch to Overwatch 2 | Details on Overwatch 2's "Defense Matrix". https://dotesports.com/overwatch/news/all-anti-toxicity-changes-from-overwatch-to-ow2
[64] 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 (July 2018), 197–213. https://doi.org/10.1016/j.cose.2018.02.016
[65] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity Detection in Multiplayer Online Games. In *Proceedings of the 2015 International Workshop on Network and Systems Support for Games (NetGames '15)*. IEEE Press.
[66] Maury Nation, Cindy Crusto, Abraham Wandersman, Karol L. Kumpfer, Diana Seybolt, Erin Morrissey-Kane, and Katrina Davino. 2003. What works in prevention: Principles of effective prevention programs. *American Psychologist* 58, 6-7 (June 2003), 449–456. https://doi.org/10.1037/0003-066X.58.6-7.449
[67] Joaquim A. M. Neto and Karin Becker. 2018. Relating conversational topics and toxic behavior effects in a MOBA game. *ENTERTAINMENT COMPUTING* 26 (May 2018), 10–29. https://doi.org/10.1016/j.entcom.2017.12.004
[68] Joaquim A. M. 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 (WI '17)*. Association for Computing Machinery, New York, NY, USA, 26–33. https://doi.org/10.1145/3106426.3106452
[69] Call of Duty. 2022. AN UPDATE, CALL OF DUTY ANTI-TOXICITY PROGRESS REPORT. https://www.callofduty.com/blog/2021/05/ANTI-TOXICITY-PROGRESS-REPORT
[70] Kyle Orland. 2015. Riot rolls out automated, instant bans for League of Legends trolls. https://arstechnica.com/gaming/2015/05/riot-rolls-out-automated-instant-bans-for-league-of-legends-trolls/
[71] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* (March 2021), n71. https://doi.org/10.1136/bmj.n71
[72] Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linares. 2019. Conversational Networks for Automatic Online Moderation. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS* 6, 1 (Feb. 2019), 38–55. https://doi.org/10.1109/TCSS.2018.2887240
[73] Craig Pearson. 2021. Dota 2's player-powered anti-griefing system is here. *Rock, Paper, Shotgun* (Jan. 2021). https://www.rockpapershotgun.com/dota-2s-player-powered-anti-griefing-system-has-just-gone-live
[74] C. Platzer. 2011. Sequence-based bot detection in massive multiplayer online games. In *2011 8th International Conference on Information, Communications & Signal Processing*. 1–5. https://doi.org/10.1109/ICICS.2011.6174239
[75] J. Prather, R. Nix, and R. Jessup. 2017. Trust management for cheating detection in distributed massively multiplayer online games. In *Annual Workshop on Network and Systems Support for Games*. IEEE Computer Society, 40–42. https://doi.org/10.1109/NetGames.2017.7991547
[76] Elizabeth Reid, Regan L. Mandryk, Nicole A. Beres, Madison Klarkowski, and Julian Frommel. 2022. Feeling Good and In Control: In-Game Tools to Support Targets of Toxicity. *Proc. ACM Hum.-Comput. Interact.* 6, CHI PLAY (Oct.

2022). https://doi.org/10.1145/3549498
[77] Elizabeth Reid, Regan L Mandryk, Nicole A Beres, Madison Klarkowski, and Julian Frommel. 2022. "Bad Vibrations": Sensing Toxicity From In-Game Audio Features. *IEEE Transactions on Games* 14, 4 (Dec. 2022), 558–568. https://doi.org/10.1109/TG.2022.3176849
[78] Valentin Riemer, Julian Frommel, Georg Layher, Heiko Neumann, and Claudia Schrader. 2017. Identifying features of bodily expression as indicators of emotional experience during multimedia learning. *Frontiers in psychology* (2017), 1303.
[79] Sarah Roberts, Stacy Wood, and Yvonne Eadon. 2023. "We Care About the Internet; We Care About Everything" Understanding Social Media Content Moderators' Mental Models and Support Needs. https://doi.org/10.24251/HICSS.2023.252
[80] Kelly Schmidt. 2023. The Results of World of Warcraft's Social Contract, One Year Later. https://gamerant.com/world-of-warcraft-shadowlands-social-contract-player-behavior-results-good-bad/ Section: Games.
[81] Dennis Scimeca. 2013. Using science to reform toxic player behavior in League of Legends. https://arstechnica.com/gaming/2013/05/using-science-to-reform-toxic-player-behavior-in-league-of-legends/
[82] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–28. https://doi.org/10.1145/3415178
[83] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 111–125. https://doi.org/10.1145/2998181.2998277
[84] Sercan Sengün, Joni Salminen, Soon-gyo Jung, Peter Mawhorter, and Bernard J. Jansen. 2019. Analyzing Hate Speech Toward Players from the MENA in League of Legends. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3312924
[85] Sercan Sengün, Joni Salminen, Peter Mawhorter, Soon-gyo Jung, and Bernard Jansen. 2019. Exploring the Relationship Between Game Content and Culture-based Toxicity: A Case Study of League of Legends and MENA Players. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. ACM, Hof Germany, 87–95. https://doi.org/10.1145/3342220.3343652
[86] L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, and the PRISMA-P Group. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 349, jan02 1 (Jan. 2015), g7647–g7647. https://doi.org/10.1136/bmj.g7647
[87] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. 2022. Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*. Association for Computing Machinery, New York, NY, USA, 2659–2673. https://doi.org/10.1145/3548606.3560599
[88] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry* 49, 4 (April 2008), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x
[89] Henry Stenhouse. 2020. CS:GO will now auto-mute abusive chatters. https://ag.hyperxgaming.com/article/9447/csgo-will-now-auto-mute-abusive-chatters
[90] 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 *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021)*. Association for Computing Machinery, New York, NY, USA, 117–118. https://doi.org/10.1145/3453892.3464897
[91] Facepunch Studios. 2013. *Rust*. Game [PC].
[92] John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3 (June 2004), 321–326. https://doi.org/10.1089/1094931041291295
[93] Paige L. Sweet. 2020. Who Knows? Reflexivity in Feminist Standpoint Theory and Bourdieu. *Gender & Society* 34, 6 (Nov. 2020), 922–950. https://doi.org/10.1177/0891243220966600
[94] Anne Clara Tally, Yu Ra Kim, Katreen Boustani, and Christena Nippert-Eng. 2021. Protect and Project: Names, Privacy, and the Boundary Negotiations of Online Video Game Players. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021). https://doi.org/10.1145/3449233
[95] Ibrahim Tanrikulu. 2017. Cyberbullying prevention and intervention programs in schools: A systematic review. *School Psychology International* (Dec. 2017), 014303431774572. https://doi.org/10.1177/0143034317745721
[96] Sian Tomkinson and Benn van den Ende. 2022. 'Thank you for your compliance': Overwatch as a Disciplinary System. *GAMES AND CULTURE* 17, 2 (March 2022), 198–218. https://doi.org/10.1177/15554120211026257
[97] 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. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376191

[98] Uel, Feak Steve, Neichus, and IceFrog. 2003. Defence of the Ancients (DotA).

[99] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence* 3, 2 (Feb. 2021), 125–133. https://doi.org/10.1038/s42256-020-00287-7

[100] Max van Haastrecht, Injy Sarhan, Bilge Yigit Ozkan, Matthieu Brinkhuis, and Marco Spruit. 2021. SYMBALS: A Systematic Review Methodology Blending Active Learning and Snowballing. *Frontiers in Research Metrics and Analytics* 6 (May 2021). https://doi.org/10.3389/frma.2021.685591

[101] C.J. Van Rijsbergen. 1977. A THEORETICAL BASIS FOR THE USE OF CO-OCCURRENCE DATA IN INFORMATION RETRIEVAL. *Journal of Documentation* 33, 2 (Feb. 1977), 106–119. https://doi.org/10.1108/eb026637

[102] Hanh Hong-Phuc Vo, Hieu Trung Tran, and Son T Luu. 2021. Automatically Detecting Cyberbullying Comments on Online Game Forums. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*. 1–5. https://doi.org/10.1109/RIVF51545.2021.9642116

[103] 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* (, Stratford, ON, Canada,) *(CHI PLAY Companion '23)*. Association for Computing Machinery, New York, NY, USA, 3–9. https://doi.org/10.1145/3573382.3616068

[104] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[105] Jiyoung Woo, Sung Hee Park, and Huy Kang Kim. 2022. Profane or Not: Improving Korean Profane Detection using Deep Learning. *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS* 16, 1 (Jan. 2022), 305–318. https://doi.org/10.3837/tiis.2022.01.017

[106] 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.

[107] Bo Sophia Xiao. 2013. Cyber-Bullying Among University Students: An Empirical Investigation from the Social Cognitive Perspective. 8, 1 (2013).

[108] Ryan Yen, Li Feng, Brinda Mehra, Ching Christie Pang, Siying Hu, and Zhicong Lu. 2023. StoryChat: Designing a Narrative-Based Viewer Participation Tool for Live Streaming Chatrooms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3580912

[109] M. Yousefi and D. Emmanouilidou. 2021. Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network. In *2021 29th European Signal Processing Conference (EUSIPCO)*. 11–15. https://doi.org/10.23919/EUSIPCO54536.2021.9616001

[110] Zhe Yu, Jeffrey C. Carver, Gregg Rothermel, and Tim Menzies. 2022. Assessing expert system-assisted literature reviews with a case study. *Expert Systems with Applications* 200 (Aug. 2022), 116958. https://doi.org/10.1016/j.eswa.2022.116958

Received February 2024; revised June 2024; accepted July 2024

Table 1. All intervention systems in our study including IDs, the respective papers, and a short description.

ID	Paper Title	Short Description of Intervention Method
1	Reid et al. [76]	Positive Voice Lines - Supportive message from virtual characters.
2	Reid et al. [76]	Riddikulus - Adding humorous messages to the text chat.
3	Reid et al. [76]	Blocklist - Control over perpetrators through muting and blocking.
4	Reid et al. [76]	Eyebleach Pictures - Generate positivity and mood relief through exposure to cute animal pictures.
5	Reid et al. [76]	Friendly Messages - Supportive message from other players.
6	Reid et al. [76]	Send a Message - Relieve the burden of dealing with a toxic player by mediation through the game.
7	Reid et al. [76]	Report misbehaving players.
8	Kou and Gui [52]	Flagging - generated by either users or automated tools and sent to an adjudication process that determines whether the fagged behavior constitutes a violation
9	Canossa et al. [13]	Prediction if a player is toxic by observing in-game behavior.
10	Murnion et al. [64]	Detection of cyberbullying through semantic analysis.
11	Märtens et al. [65]	Toxic message detection in chat logs.
12	Kou [50]	Assessment of the permanent ban as a punishment strategy.
13	Blackburn and Kwak [8]	System that predicts the outcomes of a crowd-sourced anti toxic behavior tool.
14	Kaiser and Feng [45]	Distributed player reputation system which enables players to avoid anti-social peers.
15	Neto et al. [68]	Detecting toxic conversation through chat topic recognition.
16	Neto and Becker [67]	Detecting toxic conversation through chat topic recognition.
17	Reid et al. [77]	Detection of toxicity through in-game audio fragments.
18	Ekiciler et al. [20]	Text analysis with text mining to detect sexism.
19	Balci and Salah [5]	Validation of user reports using machine learning.
20	Kou and Gui [51]	Governance system for an online game with humans and with AI.
21	Frommel et al. [29]	Evaluation technique for the quality of social interactions in video games.
22	Vo et al. [102]	Prediction method for cyberbullying in online game forums of League of Legends and World of Warcraft.
23	Stepanova et al. [90]	Verbal aggression detection and reflection for gamers.
24	Cornel et al. [17]	Cyberbullying detection using convolutional neural network.
25	Papegnies et al. [72]	Automatic abuse detection using conversational networks.
26	Yousefi and Emmanouilidou [109]	Convolutional neural network used to analyse audio samples in order to detect toxicity.
27	Lee et al. [58]	A multi stage detection system for abusive text.
28	Woo et al. [105]	Detection of profane language in Korean graphemes.
29	Busch et al. [11]	Legal tools to govern communities in World of Warcraft.
30	Andrigueto and Araujo [4]	Model of fuzzy logic that can assess aggressiveness.
31	Si et al. [87]	This paper proposes a tool that tests AI-driven text generators for toxic responses.
32	Tomkinson and van den Ende [96]	Assessment of the Overwatch endorsement system.
33	Prather et al. [75]	Decentralized algorithm for detecting cheating and griefing.
34	Brewer et al. [10]	Digital social awareness campaign through Twitch.
35	Fox and Tang [26]	Coping strategies by creative use of existing in game systems like avatars and usernames.
36	Daily [18]	Word filtering for Sony's world wide network for Playstation 2

Table 2. Themes and associated systems for types of toxicity (identifiers in bold occur in multiple categories)

Theme	Sub-theme	Description	Intervention ID
Targeted Toxic Behavior	Harassment	Behavior that repeatedly demeans, humiliates, and intimidates another player.	**9**[13], 10[64], **11**[65], **17**[77], **19**[5], 22[102], 24[17], **25**[72], **29**[11], **32**[96], **34**[10]
	Identity-Based Attacks	Toxic behavior aimed at a person's identity	**15**[68], **16**[67], 18[20], **19**[5], **25**[72], 27[58], **29**[11], **34**[10], 35[26]
General Toxic Behavior	Toxic Communication	Untargeted but undesired behavior through communication channels	**9**[13], **11**[65], **15**[68], **16**[67], **17**[77], **20**[51], 23[90], 26[109], 28[105], **30**[4], 31[87], **32**[96], 36[18]
	Non-Communication	Untargeted but undesired behavior through non-communication channels such as gameplay actions	**9**[13], **20**[51], **30**[4], **32**[96], 33[75]
	Unspecified	Intervention systems that do not specify or are not restricted to a specific type of toxicity	1[76], 2[76], 3[76], 4[76], 5[76], 6[76], 7[76], 8[52], 12[50], 13[8], 14[45], 21[29]

Table 3. Themes and associated systems for intervention sites (identifiers in bold occur in multiple categories)

Theme	Sub-theme	Group	Description	Intervention ID
Player	Exposure	-	Intervention systems that reduce exposure for the player	**2**[76], **3**[76] 35[26],
	Awareness	-	Intervention systems that create awareness amongst players	**6**[76], 14[45], 23[90], 29[11], 32[96], 34[10]
	Mood	-	Intervention systems that repair the player's mood	**2**[76], **3**[76], 4[76],5[76], **6**[76], 7[76], 8[52], 12[50], 21[29]
Platform	Human Moderators	-	Intervention systems that aid human moderators	9[13], 19[5]
	Toxic Content	Chat	Intervention systems that act on the chat channel	10[64], 11[65], 15[68], 16[67], 18[20], 22[102], 24[17], 25[72], 7[58], 28[105], 31[87], **36**[18]
	Toxic Content	Voice	Intervention systems that act on the voice channel	17[77], **21**[29], 26[109]
	Toxic Content	Non-Communication	Intervention systems that act on in-game actions.	13[8], **21**[29], 30[4]
	Toxic Content	Players	Intervention systems that influence players directly, e.g., preventing profanity in usernames	12[50], 20[51], 33[75], **36**[18]

Table 4. Themes and associated systems for intervention methods (identifiers in bold occur in multiple categories)

Theme	Description	Intervention ID
Empowering	Providing tools for players to combat the effects of toxicity themselves	3[76], 7[76], **8**[52]. 14[45], 35[26]
Supporting	Providing support to the player exposed to toxicity	1[76], 2[76], 4[76], 5[76], 6[76]
Priming	Creating awareness of what is allowed or acceptable and what is not	**23**[90], 29[11], 32[96], 34[10]
Sanctioning	Applying sanctions to players, often consisting of various types of exclusion	**8**[52], 12[50], 20[51], **23**[90], 33[75], 36[18]
Detecting	Detecting toxic behavior in different channels	9[13], 10[64], 11[65], 13[8], 15[68], 16[67], 17[77], 18[20], 19[5], 21[29], 22[102], 24[17], 25[72], 26[109], 27[58], 28[105], 30[4], 31[87]

Table 5. Themes and associated systems for evaluation approaches (Identifiers in bold occur in multiple categories)

Theme	Sub-theme	Description	Intervention ID
Model Performance	-	Intervention systems in this theme are evaluated using model performance metrics (e.g., accuracy or F1 Score).	9[13], 10[64], 11[65], 13[8], 15[68], 16[67], 17[77], 18[20], 19[5], 21[29], 22[102], 24[17], 25[72], 26[109], 27[58], 28[105], 31[87]
User Studies	Qualitative Methods	Intervention systems in this theme are evaluated using qualitative research methods like interviews or observation.	1[76], 2[76], **3**[76], **4**[76], **5**[76], **6**[76], 8[52], 12[50], 20[51], 35[26]
	Quantitative Methods	Intervention systems in this theme are evaluated using quantitative research methods like validated scales.	**3**[76], **4**[76], **5**[76], **6**[76], 7[76], 34[10], 35[26]
Simulations	-	Intervention systems in this theme are evaluated using simulations, e.g., simulating a game environment to test effectiveness.	14[45], 33[75]
Content and System Analysis	-	Intervention systems in this theme are evaluated using content or system analyses, observing and/or comparing existing interventions.	29[11], 32[96]
No Evaluation	-	Intervention systems in this theme are not evaluated in the reviewed paper.	23[90], 30[4], 36[18]

Table 6. Coding of evaluations for intervention systems in our dataset.

	evaluated with players	**not evaluated with players**	**total**
based on commercial settings	5	17	22
not based on commercial settings	7	7	14
total	12	24	36

Table 7. Cross tabulation between Intervention Methods (columns) and Types of Toxicity (rows)

	Empowering	Supporting	Priming	Sanctioning	Detecting
Harassment	0	0	3	0	8
Identity Based Attacks	1	0	2	0	6
Toxic Communication	0	0	2	3	9
Non Communication Toxic Behavior	0	0	1	2	2
Unspecified	3	6	0	2	2

Table 8. Cross tabulation between Intervention Methods (columns) and Intervention Sites (rows)

	Empowering	Supporting	Priming	Sanctioning	Detecting
Player Exposure	2	1	0	0	1
Player Awareness	1	1	4	1	0
Player Mood	3	4	0	2	0
Platform - Human Moderators	0	0	0	0	2
Platform - Content - Chat	0	0	0	1	11
Platform - Content - Voice	0	0	0	0	3
Platform - Content - Non Communication	0	0	0	0	3
Platform - Players	0	0	0	4	0

Table 9. Cross tabulation between Types of Toxicity (columns) and Intervention Sites (rows)

	Harassment	Identity Based Attacks	Toxic Communication	Non Comm Toxic Behavior	Unspecified
Player Exposure	0	1	0	0	2
Player Awareness	3	2	2	1	2
Player Mood	0	0	0	0	9
Platform - Human Moderators	2	1	1	1	0
Platform - Content - Chat	5	5	6	0	0
Platform - Content - Voice	1	0	2	0	1
Platform - Content - Non Communication	0	0	1	1	2
Platform - Players	0	0	2	2	1