Help, My Game Is Toxic! First Insights from a Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games

Michel Wijkstra, Katja Rogers, Regan L. Mandryk, Remco C. Veltkamp, Julian Frommel

ABSTRACT
Toxicity is a common problem in online games. Players regularly experience negative, hateful, or inappropriate behavior during gameplay. Intervention systems can help combat toxicity but are not widely available and even comprehensively studied regarding their approaches and effectiveness. To assess the current state of toxicity intervention research, we are conducting a systematic literature review about intervention methods for toxic behaviors in online video games. In this work-in-progress, we report the research protocol for this review and the results from a preliminary analysis. We collected 1176 works from 4 digital libraries and performed abstract and full-text screening, resulting in 30 relevant papers containing 36 intervention systems. By analyzing these intervention systems, we found: 1) Most research proposes novel approaches \( (n = 28) \) instead of analyzing existing interventions. 2) Most systems intervene only after toxicity occurs \( (n = 31) \) with few interventions that act before toxicity. 3) Only few interventions are evaluated with players and in commercial settings \( (n = 5) \), highlighting the potential for more research with higher external validity. In our ongoing work, we are conducting an in-depth analysis of the interventions providing insights into their approaches and effectiveness. This work is the first step toward effective toxicity interventions that can mitigate harm to players.

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.

KEYWORDS
toxicity, interventions, systematic literature review, online games
provide insights into their aims, which behaviors they combat, their approaches, or their effectiveness. This lack of distilled information makes it difficult to translate findings into practice (e.g., which approaches to implement in a game), assess the state of interventions (e.g., which interventions are promising), and progress the field (e.g., which novel interventions to develop).

To close this gap, we are conducting a systematic review of prior work that informs future intervention systems for toxic behaviors, aligned with the PRISMA-P standard [41, 47] and pre-registered on OSF. In this work-in-progress, we report findings from a preliminary analysis of the ongoing review. Using a systematic database search, we have finished collecting literature on papers that could address toxicity through intervention systems. After abstract screening, full-text screening, and several rounds of discussions, we identified 36 intervention systems in 30 papers. Through iterative coding, we categorized approaches based on three characteristics.

In our analysis, we found that toxicity intervention literature overwhelmingly proposes novel (n = 28) instead of analyzing existing (n = 8) approaches. Most systems act only after toxicity has occurred (n = 31) in contrast to few interventions that help before toxicity (n = 5). Regarding the evaluation of these approaches, we observed large gaps, because only a few interventions are evaluated with players (n = 12) and even fewer are evaluated with players in commercial settings (n = 5). Overall, these results show that there are large gaps regarding the design of toxicity interventions and their evaluation. As the next steps of our research, we will continue with the review and conduct an in-depth analysis of the proposed approaches, e.g., categorizing them based on their goals, methods, and effectiveness, developing a design space for interventions and ultimately informing which approaches to pursue. In summary, we present first insights into the current state of toxicity intervention research and highlight potential research gaps. This is the first step of a work-in-progress that enables the design and development of better interventions and ultimately decreases toxicity in online games.

2 BACKGROUND

Toxicity is an umbrella term used to describe a collection of negative behaviors [54]. It is difficult to precisely define what toxicity is, as the context of the behaviors plays a large role in its perception. Where a certain type of behavior could be considered negative in the context of one game, it could be considered normal or even an important gameplay element in another. For example, in the game Rust, it is a core gameplay element to kill other players on the server, raid their houses and destroy their belongings. All these actions are inappropriate behavior, dark participation, and abusive behavior, harmful behaviors [2, 26, 31, 54]. 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 [6], they are often not intentionally harmful to users. We used this scope to guide our search to collect a broad sample of toxicity research covering different aspects of behavior that intentionally harm other players and accordingly used them to define the keywords. The keywords are the names of the behaviors we chose to include, modified to be searchable in their different tenses and variations by applying wildcards. We have 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*).
We selected four electronic libraries that are commonly used in HCI research: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. Our search was limited to a date range from 1990 up to and including 2022. These collections returned an initial set of 1906 records. By importing this collection into the Zotero Reference Manager, 730 duplicates were identified before the screening. We observed that the ACM Digital Library provided us with results that did not meet our search criteria. While our search query was limited to the abstracts of the papers, in certain cases results 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. Our intervention collection initially contained one such item (Tally et al. [52]), which did not have one of our toxicity keywords in the abstract. We initially contained one such item (Tally et al. [52]), which did not 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.

### 3.2 Database Search

We selected four electronic libraries that are commonly used in HCI research: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. Our search was limited to a date range from 1990 up to and including 2022. These collections returned an initial set of 1906 records. By importing this collection into the Zotero Reference Manager, 730 duplicates were identified before the screening. We observed that the ACM Digital Library provided us with results that did not meet our search criteria. While our search query was limited to the abstracts of the papers, in certain cases results 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. Our intervention collection initially contained one such item (Tally et al. [52]), which did not have one of our toxicity keywords in the abstract. We initially contained one such item (Tally et al. [52]), which did not 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.

### 3.3 Abstract Screening

After deduplication, we had 1176 papers left that we included in the 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, 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 and works that do not include a digital component or explicitly focus on toxic behavior (as defined by our terms). When we were unable to make a decision based on the abstract alone, we also checked the full text (see also pre-registration for criteria).

We used tool-assisted screening with active learning techniques using ASReview [55], which is an active learning tool aimed at helping authors throughout the title and abstract screening phase of systematic literature reviews. Such an approach reorders the set of items to review, 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. Two authors 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. With this, we had 20 labeled papers containing 8 relevant papers and 12 irrelevant papers, which were imported into ASReview, providing broad coverage.

The first author then screened the remaining abstracts until the previously defined stopping criteria were met. Our pre-defined stopping criteria stated that we stop 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 (≈ 118). This was reached after screening 488 abstracts, at which point we then stopped screening. The resulting collection consisted of 38 papers.

### 3.4 Full-text Screening

The 38 papers selected during our abstract screening were subjected to full-text screening. During this phase, eligibility was verified once
more. This resulted in 8 more papers being excluded for not meeting the inclusion/exclusion criteria defined in our pre-registration document. The most common exclusion reasons were: 1) not providing a concrete tool, e.g., [20, 35] and 2) not being focused on toxicity, e.g., [21, 23]. This left us with 30 relevant papers. From these papers, we extracted information using our data extraction sheet (see supplementary material) identifying separate intervention approaches in a single paper (e.g., six interventions in [45]). Through this process, we identified 36 unique intervention systems that we further analyzed. Table 1 displays the entire process and results across the steps.

3.5 Preliminary Analysis
We conducted a preliminary analysis to answer three different questions: 1) Do authors evaluate existing or novel interventions? 2) Are interventions applied before or after harm from toxic behavior occurs? and 3) Do authors evaluate their interventions in realistic settings?

The first author went through an initial coding process, assigning codes for categorizing the intervention systems. At this stage, this process has been finalized for the data required to answer the research questions presented in this analysis. These codes were discussed with the first and last author to refine the code book and discuss ambiguous cases. After a second round of coding and discussion, the results were finalized to answer our research questions.

To assess 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., [28, 53]—or as novel—if the intervention was proposed as a novel system as part of the paper, e.g., [17, 50]. 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 [9], 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 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 [4, 15]). 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 analysis of classification accuracy, e.g., [8, 42].

We coded an intervention as based on commercial settings if it was created with or applied to data from a commercial game or platform, e.g., a League of Legends dataset [10]. We applied the code not based on commercial settings if the intervention method has not been applied to such a setting, e.g., instead in a custom game designed for the experiment.

4 RESULTS
In this section, we discuss our findings.

4.1 Do Authors Evaluate Existing or Novel Interventions?
We found 28 interventions that propose a novel way of addressing toxic behavior. The remaining 8 intervention systems assess an existing intervention system. Thus, the majority of work proposes novel interventions (e.g., a new implementation of fuzzy logic [1] or a system that predicts toxicity based on gameplay actions [7]) while existing systems (e.g., the endorsement system in Overwatch [53]) are less commonly studied. On one hand, this is great, because there is a lot of value to 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 [56]. On the other hand, it is surprising that so few papers study existing interventions. Such work can provide valuable insights, such as highlighting that reporting is often misused [30], while also arguing the potential benefits like mood repair through the act of reporting [45]. Especially considering external validity, this points to a gap in the literature about the evaluation of already existing interventions.

4.2 Are Interventions Applied Before or After Harm From Toxic Behavior Occurs?
Within our collection, 31 intervention systems act after toxicity occurs. The remaining five perform this task before. We observe that the majority of intervention systems take action after toxicity, which is easiest because there is a clear trigger for intervention. These interventions include AI-based systems that detect if toxicity has occurred (e.g. [34, 37]) and the systems that provide mood repair after exposure to toxicity [45]. There are only five interventions coded as before, e.g., including work by Busch et al. [5] who describe legal agreements that aim to nudge players towards better behavior before they are toxic. Similarly, Fox and Tang [15] describe the act of gender masking as a coping strategy that women use against harassment. We included this as a potential before toxicity intervention strategy because it can help avoid exposure but needs to be facilitated through existing in-game systems like avatar and nickname selection. This is also interesting because gender masking is performed before toxic behavior occurs, however, it is likely a
We have coded the evaluation of each intervention system on two factors: 1) Is the intervention system based on a commercial setting and 2) has the intervention system 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., [12, 33]). The remaining 14 interventions are not based on commercial settings (e.g., [43, 50]). Out of the 36 intervention systems, 12 are evaluated with players (e.g., [29, 45]) while the majority of the intervention systems (n = 24) was not evaluated with players (e.g., [34, 48]). Bringing this together, only five of the interventions are evaluated with players and based on commercial settings ([4, 15, 28–30]).

A full overview of our findings is reported in Table 2. While there is a lot of value to other evaluation approaches, it is ultimately necessary to evaluate a toxicity intervention system in a commercial setting using real players, providing insights about external validity. In the current state of literature, there seems to be a gap between academic research about interventions and their evaluation in commercial settings and with players.

5 DISCUSSION
While still a work-in-progress, our preliminary results provide several insights into the state of toxicity intervention literature. We observe that the overwhelming majority of intervention systems in literature are novel. On one hand, this is great, because there are increasingly more new approaches to dealing with the complex issue of toxicity. On the other hand, we see value in more assessments of existing systems to better understand what strategies are effective. As such, we argue for more future work that assesses toxicity interventions that are commonly used in games.

Most of the intervention systems in our study intervene after harm has already been done. In contrast, there are only a few approaches that act before toxicity. Such interventions have huge potential because they prevent harm instead of mitigating it. We understand that this is challenging because it relies on approaches that reliably detect toxicity to trigger intervention, which is hindered by the lack of tools and subjectivity of toxicity [16], or the development of systems that lead to more positive communities and player interactions [33], which is difficult due to increasing normalization of toxicity [2]. We highlight a gap in harm-preventing systems and hope that future research further explores these approaches because of their potential value.

We have also observed a gap in the evaluation of interventions with potential for strengthening their external validity. Only five interventions are evaluated with players and based on commercial settings. Such evaluations are beneficial because of the dynamics of play. Evaluation of interventions in controlled settings is important, but not necessarily representative of how they would work in existing commercial games that are subject to other expectations and norms. Similarly, it is important to bring players into the evaluation to assess effectiveness and user acceptance. For example, study participants considered some mood repair approaches as silly, highlighting the need for appropriate integration and subsequent evaluation [45]. Thus, we consider work especially valuable if interventions are evaluated with players in commercial settings. These papers include evaluations of existing systems in League of Legends [28–30] and coping strategies that we could facilitate through systems [15]. Only Brewer et al. [4] 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 our small collection of five intervention systems that act before toxic behavior affects another player, three are evaluations of existing systems [5, 15, 53] and the fourth work by Daily [9] dates back to 2006. Since then, the paper by Brewer et al. [4] is the only one in our dataset proposing an approach for a harm-preventing intervention system. To summarize, we think there is a lot of potential for more and stronger evaluation of toxicity interventions in commercial settings and with players.

5.1 Limitations
First, our methodology for finding and screening literature does not reach every possible element of the problem we are researching. For example, work by Grace et al. [18] describing code of conduct agreements for online games is not found by our search strategy, because the paper does not specifically mention toxicity in its title or abstract, even though it can help combat toxicity. As such, not all relevant work can be included in the systematic review to maintain a manageable scope. Second, the results reported in this work-in-progress are based on descriptive statistics and preliminary analysis. The results of the full review will provide further insights as these will include a full analysis of the data. Third, our review is limited to toxicity interventions in multiplayer games, in which toxicity happens. With this, we ended up excluding work like Komac et al. [25] that presents an interesting approach aimed at increasing awareness about trolling through serious games. This can help combat toxicity but is not a component of the game, in which toxicity occurs, leading us to exclude it. However, such approaches could be applied in multiplayer games where they can combat toxicity. Fourth, we used active-learning approaches in screening, which is a novel approach that still requires more validation and comparison to traditional reviewing approaches. Lastly, we recognize that we only review academic work and cannot make claims about the progress made in the games industry.

5.2 Next Steps
After presenting these results, we will continue with the next steps of the systematic literature review. This step will include processing all the remaining data in our collection. We have collected a
significant amount of qualitative data, which we will analyze using thematic analyses. Our research will attempt to identify various more properties of intervention systems, for which there is limited prior work. We will use a combination of inductive and deductive approaches, using existing work to guide analysis when available (e.g., HCI contributions [56]). With this, we will answer the following research questions: 1) What intervention methods for toxic behaviors in online video games currently exist? 2) How can intervention methods help online video games combat toxicity? Each of these research questions will be supported by sub-questions. These are available in our protocol pre-registration document. By answering these research questions, we aim to provide a starting point for future research into toxicity intervention methods. Our work will allow researchers to progress the field by enabling them to build on each others’ work.

6 CONCLUSION
Our initial analysis allowed us to make the following contributions: We found that intervention systems proposed by academic literature are mostly novel, while few works explore the functionality of existing systems. We also found that there are far fewer intervention systems that act before toxicity affects another player than interventions acting after toxicity, pointing to a gap in approaches of existing systems. We also found that there are far fewer intervention methods helping online video games combat toxicity? Each of these research questions will be supported by sub-questions. These are available in our protocol pre-registration document. By answering these research questions, we aim to provide a starting point for future research into toxicity intervention methods. Our work will allow researchers to progress the field by enabling them to build on each others’ work.

REFERENCES
[7] A. Canossa, D. Salimov, A. Azadvar, C. Harteveld, and G. Yannakakis. 2021. Forging a partnership: Using existing work to guide analysis when available (e.g., HCI contributions [56]). With this, we will answer the following research questions: 1) What intervention methods for toxic behaviors in online video games currently exist? 2) How can intervention methods help online video games combat toxicity? Each of these research questions will be supported by sub-questions. These are available in our protocol pre-registration document. By answering these research questions, we aim to provide a starting point for future research into toxicity intervention methods. Our work will allow researchers to progress the field by enabling them to build on each others’ work.