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

Michel Wijkstra
m.wijkstra@uu.nl
Utrecht University
Utrecht, 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
r.c.veltkamp@uu.nl
Utrecht University
Utrecht, Utrecht, Netherlands

Julian Frommel
j.frommel@uu.nl
Utrecht University
Utrecht, Utrecht, Netherlands

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 or 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**.

CHI PLAY Companion '23, October 10–13, 2023, Stratford, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0029-3/23/10…$15.00
https://doi.org/10.1145/3573382.3616068

KEYWORDS

toxicity, interventions, systematic literature review, online games

ACM Reference Format:
Michel Wijkstra, Katja Rogers, Regan L. Mandryk, Remco C. Veltkamp, and Julian Frommel. 2023. Help, My Game Is Toxic! First Insights from a Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion '23), October 10–13, 2023, Stratford, ON, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3573382.3616068

1 INTRODUCTION

Toxicity is a problem that players and developers of most multiplayer games have to deal with [32, 33], with research going back to Dibbell's work in 1994 [11]. While game developers, researchers, and players recognize the problem and have started combating toxicity, it has not been solved but instead worsened. A critical report from the Anti-Defamation League in 2022 [33] revealed that five out of six adults (86%) have experienced harassment in online play. This is a worrying statistic, as such toxic actions disrupt the players' enjoyment and performance [45].

One approach for combating toxicity and its effects are *intervention* systems, which we consider any digital system component that helps combat toxicity or its effects. Intervention systems commonly assist in or attempt to completely automate the process of moderation. However, they can be more widely applied, for example, enabling players to manage their own behavior or even prevent toxicity exposure. Throughout this work we will use the terms "interventions" and "intervention systems" interchangeably. Many of these approaches are common in commercial games, 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) [36, 40, 43, 49]. Academic research has contributed to this by proposing new interventions (e.g., [7, 34, 42]) or analyzing existing approaches (e.g., [5, 15, 53]). There is evidence for the value of interventions in non-gaming contexts [14], in commercial games [39, 46], and games research [45], highlighting that such approaches could be beneficial for combating toxicity and its effects. However, there is no comprehensive overview of toxicity interventions in games that could 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 could be considered toxic in other games. Generally, behavior is considered toxic if it violates the rules and social norms of the game [2].

Toxicity has negative effects on both players and game developers. For players, common effects are loss of enjoyment [54], stress and sense of losing control [45], lower in-game performance [26, 57], and on a larger scale diminished quality of community feel due to toxicity enabling more toxicity [27]. Game developers also experience the negative effects of toxicity. Deviant behavior has a negative effect on user retention [19], which is crucial when operating an online game. Further, it creates negative associations with the game, making it harder to attract new users [22, 38]. Toxicity is often described as a result of the online disinhibition effect, which causes a lack of restraint in online communication compared to in-person communication [51]. This lack of restraint, the fast-paced nature of online video games, competitiveness, and lack of consequences result in a high probability of repeated toxic behavior [26]. Prior work [2] showed that players with higher toxic online disinhibition (and moral disengagement) also perceive behavior as less toxic, highlighting that subjectivity complicates this problem.

Interventions are one way to deal with toxicity. By designing systems or game elements that prevent toxicity or by creating reactive systems that actively monitor and enforce rules, games can mitigate harm. Various intervention methods have been proposed, addressing different types of toxic behaviors and tackling the issue from different angles. For example, work by Kou and Gui [30] investigates reporting systems, as a common way of addressing toxicity. Blackburn and Kwak [3] take a different approach and evaluate toxicity prediction, i.e., predicting if messages are toxic, which can be used for sanctioning. Another possible approach is demonstrated by Reid et al. [45], who suggest supporting the victim instead of punishing the perpetrator. These works demonstrate that many approaches currently exist within academic literature, but no work has comprehensively reviewed the state of toxicity intervention research.

3 METHODS

We are conducting a systematic literature review to assess the current state of the literature on intervention methods for toxic behaviors in online video games. We designed a review protocol based on the PRISMA-P guidelines [47] using keyword definitions and anchor papers to guide a database search in abstracts and titles. The results of this database search were deduplicated, resulting in 1176 unique papers that were screened for inclusion, with the remaining 30 papers then coded regarding their characteristics. This research protocol was pre-registered on OSF.

3.1 Toxicity Scope, Keywords, and Anchor Papers

As toxicity is used so broadly, we defined the scope for our study, in which we include the following behaviors: harassment, abuse, hate speech, insulting, griefing, trolling, offending/offensive behavior, inappropriate behavior, dark participation, and abusive behavior, following prior work that investigated different aspects of toxic and 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*`

`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)`. These wildcards allow us to capture variations of common words used to describe toxicity, e.g., `harass*` to capture the terms harass, harassed, harassment, and harasser. These queries were adapted to work for the different databases and searched based on abstracts and titles (see pre-registration for the full queries). We selected a set of 10 papers that matched the topic of toxicity during our initial experimentation with search terms. 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. [45], Kou and Gui [30], Canossa et al. [7], Murnion et al. [37], Martens et al. [38], Kou [28], Blackburn and Kwak [3], Kaiser and Wu-Chang Feng [24]. We selected these papers because they describe different approaches and goals to combat toxicity, covering different authors, fields, databases, and publication years. Two papers about toxicity were selected to test the broad query: Kowert [31] and Kordyaka et al. [26]. These should appear in the database search but 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.

3.2 Database Search

We selected four electronic libraries that are commonly used in HCI research: ACM Digital Library: The ACM Guide to Computing Literature, 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 discussed including this paper, because it was deemed a valuable contribution. However, during full-text screening, this work did not meet the inclusion criteria and was removed from the collection. To test the query and comprehensiveness of the database search, we verified that our results included the 10 anchor papers described previously.

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).

Table 1: Process of the review and results across the different steps.

Data source	Items
ACM Digital Library	372
IEEE Xplore	61
Scopus	929
Web of Science	544
Total collected	**1906**
Duplicates removed (auto & manual)	730
Eligible for abstract screening	**1176**
Irrelevant removed (ASReview)	1138
Eligible for full-text screening	**38**
Excluded (criteria not met)	8
Total papers	**30**
Intervention systems extracted	**36**

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

Table 2: 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

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 reactive action to prior experienced toxic behavior. Furthermore, we coded two approaches for toxicity detection [7, 12] as **after** toxicity, even though the papers explain that their systems could be used in a proactive matter (i.e., predicting toxicity before it happens). While this is possible, neither paper demonstrates this. Yet, such intervention systems could be considered a more desirable solution. When an intervention intervenes before toxicity affects another player, we are able to prevent any harm to them. When dealing with toxicity, harm prevention is always better than harm mitigation.

4.3 Do Authors Evaluate Their Interventions in Realistic Settings?

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., [13, 24]). The remaining 14 interventions are **not based on commercial settings** (e.g., [44, 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 [53], 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 Komaç 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 that can prevent harm rather than mitigate it. Lastly, our analysis shows that only a few validation approaches use evaluation approaches with players and based on commercial settings, pointing to potential improvements regarding external validity. These findings highlight multiple research gaps in toxicity intervention research. We provide initial insights into the state of toxicity intervention systems research and substantiate the need for a full systematic literature review, which will aid the progression of the field.

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] Joaquim Alvino de Mesquita 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

[11] 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

[12] E. Reid, R. L. Mandryk, N. A. Beres, M. Klarkowski, and J. 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

[13] 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

[14] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting Online Harassment of Women Journalists. In *Mediating Misogyny: Gender, Technology, and Harassment*, Jacqueline Ryan Vickery and Tracy Everbach (Eds.). Springer International Publishing, Cham, 311–332. https://doi.org/10.1007/978-3-319-72917-6_16

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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.

[22] Janne Huuskonen. 2022. Toxicity and the hidden dangers of shoddy moderation. https://utopiaanalytics.com/toxicity-and-the-hidden-dangers-of-shoddy-content-moderation/

[23] 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

[24] 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.

[25] 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

[26] 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.

[27] 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

[28] 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

[29] 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

[30] 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

[31] Rachel Kowert. 2020. Dark Participation in Games. *FRONTIERS IN PSYCHOLOGY* 11 (Nov. 2020). https://doi.org/10.3389/fpsyg.2020.598947 Place: AVENUE DU TRIBUNAL FEDERAL 34, LAUSANNE, CH-1015, SWITZERLAND Publisher: FRONTIERS MEDIA SA Type: Article.

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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.

[39] Call of Duty. 2022. AN UPDATE, CALL OF DUTY ANTI-TOXICITY PROGRESS REPORT. https://www.callofduty.com/blog/2021/05/ANTI-TOXICITY-PROGRESS-REPORT

[40] 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/

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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/

[47] 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

[48] 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

[49] Henry Stenhouse. 2020. CS:GO will now auto-mute abusive chatters. https://ag.hyperxgaming.com/article/9447/csgo-will-now-auto-mute-abusive-chatters

[50] 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

[51] John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3 (June 2004), 321–326. https://doi.org/10.1089/1094931041291295

[52] 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

[53] 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

[54] 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

[55] 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

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

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

Received 2023-06-22; accepted 2023-08-03