Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming

Nicole A. Beres, Julian Frommel, Elizabeth Reid, Regan L. Mandryk, Madison Klarkowski
<firstname.lastname>@usask.ca
The Interaction Lab, University of Saskatchewan
Saskatoon, Saskatchewan, Canada

ABSTRACT

Video game toxicity, endemic to online play, represents a pervasive and complex problem. Antisocial behaviours in online play directly harm player wellbeing, enjoyment, and retention—but research has also revealed that some players normalize toxicity as an inextricable and acceptable element of the competitive video game experience. In this work, we explore perceptions of toxicity and how they are predicted by player traits, demonstrating that participants reporting a higher tendency towards *Conduct Reconstrual*, *Distorting Consequences*, *Dehumanization*, and *Toxic Online Disinhibition* perceive online game interactions as less toxic. Through a thematic analysis on willingness to report, we also demonstrate that players abstain from reporting toxic content because they view it as acceptable, typical of games, as banter, or as not their concern. We propose that these traits and themes represent contributing factors to the cyclical normalization of toxicity. These findings further highlight the multifaceted nature of toxicity in online video games.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Computer games**.

KEYWORDS

games, toxicity, toxic, normalization, moral disengagement

ACM Reference Format:
Nicole A. Beres, Julian Frommel, Elizabeth Reid, Regan L. Mandryk, Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445157

1 INTRODUCTION

Multiplayer gaming is a popular and prevalent pastime [15, 34], with social play known to provide many benefits to gamers—such as providing social support [35, 52], combating loneliness [14], and improving wellbeing [31]. Despite this, multiplayer gaming—particularly in online contexts—can also harm players by exposing them to the toxic behaviors of other players. Toxicity refers to various types of negative behaviors involving abusive communications directed towards other players (i.e., harassment, verbal abuse, and flaming) and disruptive gameplay that violates the rules and social norms of the game (i.e., griefing, spamming, and cheating) [2, 18, 29, 42, 54].

Preventing toxic behavior is a critical issue for game companies—their revenue is threatened [26] as toxicity contributes to churn (i.e., player attrition) and discourages new players from joining. To combat toxicity, game studios (e.g., Riot, Electronic Arts, Ubisoft) and companies (e.g., Twitch) formed the Fair Play Alliance coalition to encourage healthy player interactions in online gaming and build communities that are free of harassment, discrimination, and abuse [17]. Additionally, toxic behaviors are a form of cyberbullying [28] that is often directed at players from marginalized groups. Several studies have highlighted the gendered violence of harassment of women in gaming [12, 19, 27, 40], and that LGBTQ players [3] and players of color [20, 21] are disproportionately the target of harassment.

Players and game companies both have a vested interest in preventing harmful toxic play. Despite this, toxicity still persists, raising questions of why it remains so prevalent in online gaming. One potential reason is that exposure to toxic behavior is known to perpetuate it: a recent study shows that being a victim of toxic abuse in multiplayer online battle arena (MOBA) games increases a player's chance of perpetuating toxic behavior [26]. Another potential reason is an embedded belief that toxicity is an inextricable element of how gamers interact in competitive gaming contexts [2]; that is, that gamers will be gamers. This tacit acceptance of negative and abusive behaviors, justified as being simply part and parcel of the gaming context, represents a normalization of toxic behaviors within gaming culture.

Normalizing toxic behaviors in games is problematic because it creates a cycle of reciprocal perpetuation. The theory of normalized behavior [24, 39] suggests that those who are more approving of a particular behavior will be more likely to engage in that behavior, and also suggests the reverse—that engaging in a particular behavior will reinforce normative beliefs about the acceptability of it. In the context of online games, this suggests that those who engage in toxic behaviors will normalize their beliefs about toxicity, and that those with normalized beliefs will be more approving of toxic behaviors in games. Importantly, this escalating cycle of behavior–normalized beliefs–behavior means that *if games companies, players, and researchers want to combat toxicity in gaming, we have to understand how toxic behaviors become normalized in games, and which players are at risk of perpetuating this cycle.*

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445157

Some toxic behaviors are overt, hostile, and aggressive (e.g., abusive communications, flaming), and thus can be linked with attitudes about traits that might facilitate normalizing these behaviors. For example, recently, normative beliefs about cyberaggression predicted cyberaggressive behavior in an online gaming context [23]; and toxic online disinhibition (i.e., the loss of inhibition in online contexts that can lead to hostile behaviors and inappropriate behaviors) predicted these same types of toxic behaviors [26] in MOBAs. However, there are other toxic behaviors that are more passive-aggressive, subtle, or insidious. For example, insults and hate speech can be overt—but can also take a more underhanded form, in which the receiver may feel harmed but not be able to point to its definitive source; disruptive intentions (i.e., griefing) can be visible, but may also be concealed (e.g., intentionally sabotaging strategy under a veil of ignorance); or trash talk can move seamlessly from playful ribbing into hurtful ridicule and mockery, with no clear distinction as to when the transition occurs.

Similar to psychological oppression [49], gaslighting [1], and covert emotional abuse [50], these insidious forms of toxic behavior are less easily identified and called out as explicitly toxic by players, game studios, or spectators. Therefore, they may be easier to justify as simply part of the gaming context, supporting their normalization and integration into gaming culture through the escalating cycle previously described. A problem is that their insidiousness, by definition, makes these types of toxic behaviors hard to identify, difficult to connect with existing normative beliefs, and challenging to predict from individual traits that can be measured or monitored. We argue that the framework of moral disengagement [6] provides value for understanding how toxic behaviors are normalized in games, particularly for these covert and insidious behaviors that are harder to explicitly identify than their overt and aggressive counterparts. The concept of moral disengagement was introduced by Bandura [4] to help understand behaviors that deviate from personal morals and values. Used initially in the context of schools (e.g., [36]) and prisons (e.g., [45]), moral disengagement has recently been applied to explain antisocial behaviors in competitive sports contexts (e.g., [6, 56]), which are arguably similar in nature to competitive online gaming. In this paper, we use these frameworks of moral disengagement, online disinhibition, and aggression to question: *in what way are players rationalizing and justifying toxicity as acceptable behavior, and which players are most at risk of normalizing toxic behaviors in online games?*

To this end, we undertook a mixed-methods approach to study the perception of toxicity in games. Our results point to moral disengagement and toxic online disinhibition as predictors of decreased perception of toxicity (i.e., viewing toxic content as less severe). We find that participants who don't report toxic behaviors are susceptible to rationalizing them by normalizing the behaviors as appropriate, tolerable, or even beneficial within the context of online gaming. We propose that moral disengagement renders players more susceptible to the normalization of toxic online behaviours, and that the normalization of toxicity represents an impediment to reporting these behaviors. We propose that, in order to erode toxicity in online games and promote healthy communities, we must interrupt the cycle of normalization.

2 BACKGROUND

Online environments provide a plethora of phenomena that are of interest to the study of GUR and HCI for understanding player behaviors and perceptions. The nature of online environments provide a disinhibiting effect due to increased anonymity and invisibility, which afford individuals the freedom to express themselves in ways they would refrain from exhibiting in the physical world [51]. While these expressions are often harmless, they may also manifest as various forms of negative behaviors. In the context of video games, these negative behaviors are encompassed within the umbrella term of 'toxicity'. The term toxicity refers to various types of disruptive behaviors, involving both abusive communications directed towards other players, and disruptive gameplay behaviors that violate the rules and social norms of the game [2, 18, 29, 42, 54]. Although these behaviors are conventionally accepted as toxic, providing a clear definition of toxicity is complex due to differences that exist across various online games. Behaviors that are considered toxic vary across the contexts of different games and the established norms, rules, and player expectations of each community.

Behaviors commonly associated with toxicity include 'trolling', i.e., verbal or in-game behaviors intended to provoke and antagonize other players [13, 54], 'flaming', i.e., aggressive or derogatory language [29] or hostile expressions [53], 'griefing', i.e., play styles that intentionally disrupt the gaming experience of another player [18], and 'spamming', i.e., repeated disruptive use of online communications [46].

2.1 Patterns and Perceptions of Toxicity

Researchers have investigated the characteristics of toxic players to better understand the factors that contribute to in-game toxicity. Using player feedback and game metrics data from *League of Legends*, Shores et al. [44] identified several patterns of toxic players—such as playing in more competitive (e.g., ranked) game modes and playing with friends. Further, less experienced players were more susceptible to discontinuing play after encountering toxic players.

These findings are paralleled with a large dataset from the game *World of Tanks*, suggesting experienced and skillful players are more likely to commit toxic behaviors—which may be explained by continuous exposure to toxicity [42]. Supporting this, Kordyaka et al. [26] found that past victimization experiences of toxicity (i.e., the occurrence with which a player has been the target of toxicity in the past) led to higher levels of toxic disinhibition (i.e., performing negative behaviors, such as harassment). This finding suggests that being the recipient of toxic behaviors in the past leads to the acceptance of toxic behaviors, which may lead victims of toxicity to emulate such behaviors in the future [13, 26].

Studies have also sought to understand player perceptions of toxicity and motivations for engaging in toxic behaviors. Mattinen and Macey [33] examined the interplay of age and verbal abuse in *Dota 2*. It was found that older *Dota 2* players perceived communication abuse more seriously than younger players, and were also more likely to participate in such abuse. However, age was negatively correlated with placement in the low priority pool of players (i.e., a penalty applied to players that were reported for engaging in behaviors that violate the terms of the game), which

suggests that younger *Dota 2* players engage more often in toxic behaviors that would warrant a report. While these results initially seem contradictory, they may be explained by the differing perceptions of older and younger players regarding what constitutes communication abuse. The authors suggest that younger players perceive communication abuse less seriously than older players due to early exposure and normalization of toxic behaviors within online environments. These young players may in turn emulate such behavior, resulting in a cycle of toxicity whereby placement in low priority pools consisting of other toxic players further exposes young *Dota 2* players to toxic behavior.

Evidence suggests that communities play a role in asserting which behaviors become normative and accepted within the community's culture [13, 37]. In a study that investigated the motivations of self-identifying trolls, Cook et al. [13] identified that trolling is accepted within the online gaming community as an inescapable, self-perpetuating phenomenon. All of the self-confessed trolls within the sample reported being a recipient of trolling in the past and perpetuated the cycle by engaging in such behaviors themselves, thus normalizing the behavior. In a similar vein, research by Rainey and Granito [37] explored the notion of normative rules favoring trash talk (a form of abusive or belittling speech, used to establish or support social hierarchy [16]) among college athletes and the motivations for engaging in such behavior in sport. Participants reported that they engaged in trash talk to both motivate themselves, and hinder the motivation and performance of their opponents to gain a competitive advantage [37]. Athletes learned to trash talk primarily from their teammates and opponents, suggesting trash talk is a normative behavior among college athletes. Adinolf and Turkey [2] likewise found that collegiate esports club members rationalized and normalized negative behaviors as endemic to competitive gaming culture, and thus were disinclined to report these behaviors. These studies indicate that the community plays an influential role in the normalization and acceptance of toxic behaviors by modelling such behaviors to others, which in turn motivates players to emulate these behaviors themselves. While many players agree that toxicity is prevalent within online games and have admitted to engaging in toxic behaviors [13], toxicity is often not reported [28]. Under reporting may be explained by differing perceptions of what constitutes toxicity [2], reduced compulsion to report if a player is not personally harmed or affected by the toxic behavior (e.g., if they are not the target) [5], and the disinhibiting effect anonymity has on player behavior and perceptions [28, 51].

While existing research has focused on what toxicity is and why it happens, there is a lack of understanding regarding how toxicity is perceived due to a lack of consensus regarding what constitutes toxicity among various online gaming communities, and players with different demographic backgrounds and personality traits. To this end, we explore player perceptions of toxicity through the lens of online disinhibition, aggression, and moral disengagement in sport to identify patterns that may emerge regarding what constitutes toxicity in the context of video games.

2.2 The Online Disinhibition Effect

The online disinhibition effect (ODE) refers to the perceived freedom an individual feels in online environments to express themselves in ways they would refrain from exhibiting in offline settings due to decreased behavioral inhibitions [51]. This disinhibiting effect consists of two components: benign disinhibition (i.e., positive behaviors, such as acts of kindness) and toxic disinhibition (i.e., negative behaviors, such as hostile expressions). Benign disinhibition allows us to share personal feelings we would be reluctant to share otherwise, or display unusual acts of kindness and generosity. Toxic disinhibition may be expressed through hostile language, swearing, threats, and toxic behaviors.

While these components interact with each other to produce a disinhibiting effect, the extent of the effect may be influenced by personality variables. In the context of video games, studies have focused on the disinhibiting effect of anonymity and invisibility as online gaming environments facilitate high degrees of these interactions [11, 26, 29]. Previous studies suggest that online disinhibition is a predictor of toxic behaviors in online gaming environments [26, 54, 55]. While both components of online disinhibition (i.e., benign and toxic disinhibition) have been shown to contribute to the occurrence of negative behaviors online [55], recent studies suggest that toxic disinhibition is a more meaningful predictor of toxic behaviors in the context of video games [26, 54].

2.3 Moral Disengagement in Video Games

Moral disengagement, originally introduced by Bandura [4], refers to a process by which individuals rationalize engaging in immoral behaviors to avoid feelings of self-condemnation [4]. When an individual behaves in a manner that violates their morality, they attempt to validate their decision by disengaging self-sanctions from the behavior. According to Bandura [4], there are eight mechanisms individuals may employ in rationalizing their behaviors:

- Moral justification: reframing a decision to serve a moral purpose.
- Euphemistic labelling: language that detracts from the emotional intensity of the subject being referenced.
- Advantageous comparison: minimizing immoral severity by comparison to an act of greater immorality.
- Diffusion of responsibility: lack of personal accountability for action when others are present.
- Displacement of responsibility: lack of personal accountability when under the authority of another person.
- Distortion of consequences: minimization of the harm caused to others by actions.
- Dehumanization: stripping people of human qualities.
- Attribution of blame: attributing blame to others to avoid personal accountability.

In recent years, moral disengagement has been explored within the context of sports [6, 43, 56], revealing that moral disengagement has a positive association with antisocial behavior, and a negative relationship with prosocial behavior [6]. Pursuant research by Stanger et al. [48] found that moral disengagement was associated with lower anticipated guilt and higher antisocial behavior when playing a sport, and that the mechanisms of moral disengagement enable individuals to engage in antisocial behaviors by reducing

the regulatory role of guilt in managing moral behaviors. Athletes that display antisocial behaviors in sport appear to justify their behaviors by utilizing the mechanisms of moral disengagement (e.g., attribution of blame) to reduce anticipated guilt and avoid feelings of condemnation [4, 48]. It is also suggested that athletes may be less morally attentive in sport than in general life due to 'bracketed' moral reasoning [43]. This bracketing of morality may involve reducing moral attentiveness during competition, affording athletes a safeguard to engage in antisocial behaviors to acquire personal or team gain while playing a sport [8].

Recent research efforts have explored moral disengagement in online games, and position the topic as a critical avenue for further investigation. A study by Sparrow et al. [47] provides a preliminary framework of player immorality termed 'Apathetic Villager Theory', highlighting six themes (i.e., reactive morality, village reputation, masochism and schadenfreude, response paradox, threat threshold, and dark mirror) regarding player's ethical standpoints on behaviors in games. The authors suggest that further research should consider how the attitudes identified in 'Apathetic Villager Theory' interact in multiplayer games by adapting these considerations to larger-scale quantitative works and questionnaires. We adapt and apply a quantitative measure of Moral Disengagement in Sports (MDSS) [6] in the present study to further examine player attitudes and perceptions of toxicity within the multiplayer gaming context. While moral disengagement has been studied in online environments, the use of the MDSS is novel in games research.

Studies have also striven to understand the role of moral disengagement in both single-player and multi-player online gaming contexts. Research by Hartmann [22] suggests that players morally disengage in violent single-player video games by framing violence against virtual characters as an acceptable act, allowing players to enjoy the violence rather than experience guilt. In a similar vein, Klimmt et al. [25] found that players engage in moral management in the game world by adopting coping strategies that alleviate moral conflicts concerning their violent behaviors in gaming contexts. The authors conclude that moral management is absent from multiplayer combat games and applies only to single-player contexts that include narrative frameworks. However, Carter and Allison [10] found that moral management applies in *DayZ*, a survival themed multiplayer combat game wherein players are burdened with the choice of whom to kill, thus introducing moral conflicts in decision making and corresponding feelings of guilt when actions violate one's moral code. However, killing players in *DayZ* is an optional strategy in the game to secure resources for oneself, and does not address the motivations for engaging in toxic behaviors towards players in other contexts where player combat is the main objective of the game. These findings suggest further investigation concerning moral disengagement multiplayer video games is warranted.

Additional research suggests that players engage in moral disengagement in video games to justify immoral behaviors. In one such study, Shafer [41] found that, in line with the concept of bracketed morality, players who adopted moral disengagement mechanisms were more likely to make immoral choices in games. Players defended their decisions by reasoning that their actions had no real consequences, as the scenario was 'just a game'. In another study, Lee et al. [30] found that a player's moral positioning (i.e., preference for evil roles or characters) in *League of Legends* was affected by both aggressive and competitive motivation. Players with higher levels of these motivations were more likely to adopt an evil moral position regardless of the player's dispositional moral identity, and engage in more disruptive behaviors. Taken together, these findings suggest that disruptive behaviors in games are influenced by bracketed moral reasoning, moral positioning, and aggressive and competitive motivation.

While moral disengagement has been studied across various contexts, the use of the MDSS has not been applied to online gaming environments. Particularly in competitive online gaming, players might show similar characteristics as athletes in sports, hinting at the potential value of moral disengagement theory in understanding toxicity. For example, players of online games may implement mechanisms of moral disengagement to justify behaving antisocially when playing a game, which may normalize such antisocial behaviors in online gaming environments [4, 23].

2.4 Aggression in Video Games

Antisocial behaviors, such as aggression and hostile expressions, are commonly linked to online gaming [13, 23]. Hilvert-Bruce and Neill [23] examined whether players normalize aggression in online games and whether these beliefs perpetuate aggressive behaviors in games. Participants were provided a scenario describing an instance of verbal harassment occurring either online (i.e., multiplayer game) or offline (i.e., boardgame) and were asked whether they found the harassment acceptable. Harassment was perceived as normal within the online gaming scenario and normative beliefs regarding aggression predicted aggressive behaviors in games. These findings suggest that as antisocial behaviors such as aggression become normalized within online gaming, such behaviors are more likely to be perpetuated and tolerated by players with these normative beliefs. Therefore, players might normalize antisocial behaviors in games through the mechanisms of moral disengagement and their tendency to engage in antisocial behaviors, such as aggression, might further help explain why players engage in toxic behaviors.

3 PRESENT STUDY

Our goal is to explore perceptions of toxicity in online gaming and how they are predicted by player traits, such as online disinhibition, moral disengagement in games, and aggression. We also seek to understand differences in willingness to report toxic behaviour, the mechanisms that support normalization of toxicity, and the interplay between the normalization of toxicity and player traits. Examining traits may provide a better understanding of how players of differing backgrounds perceive and define toxicity. Likewise, an exploration of how players contextualize and rationalize toxicity will improve our understanding of normalized toxicity in online games. We generated four Research Questions:

RQ1. Are there traits that predict how toxic an interaction is perceived to be?

RQ2. What informs participants' decisions to report toxic behaviors?

RQ3. How do players rationalize not reporting observed toxic behaviors?

RQ4. How do traits relate to the rationalization of toxic behaviors in a gaming context?

4 STUDY METHODS

We conducted a mixed-methods analysis on data gathered in an online experiment. To investigate participant traits, we disseminated inventories querying tendency towards moral disengagement in games, online disinhibition, and aggression. Participants were then prompted to rate the perceived toxicity of social interactions from an online game, alongside items querying their inclination and reasoning for reporting or not reporting those interactions.

4.1 Procedure

The online study was deployed on Amazon Mechanical Turk (MTurk), which has been shown to be useful in behavioral research [32]. We limited respondents to those with experience playing online games, and requested familiarity with the game *Overwatch*. MTurk workers who passed these exclusion criteria were able to view the study information on the website's recruitment board. Participants were paid $12USD for their participation, which took less than one hour to complete. After providing informed consent, participants answered questionnaires about their demographics, gaming preferences and history, as well as trait inventories concerning moral disengagement in games, online disinhibition, and aggression. Participants were then asked to listen to the full duration of three audio clips (presented in random order) sampling social interactions from *Overwatch*, and were prompted after each clip to rate perceived toxicity. Finally, participants stated whether they would report the behaviours if they were in that match ('Yes', 'No', 'Unsure'), and were prompted to explain the rationale behind their decision to report or not report. Following completion of the experiment, participants were debriefed, thanked, and directed to the remuneration information. The study was conducted under ethical approval received from the Behavioral Research Ethics Board of the University of Saskatchewan.

4.2 Game and Clip Selection

Overwatch is an online multiplayer first-person shooter developed by Blizzard Entertainment. Players are assigned into two teams of six, and play a hero in either a damaging, defensive, and supportive role. In *Overwatch*, players work together as a team to secure and defend points or escort a payload across the map. The average match length ranges from 15-25 minutes, with each round lasting between 5-12 minutes. The game was selected due to its integrated voice-chat feature (enabling increased opportunities for toxic interactions amongst players) and for its relatively short game duration, reducing the potential for participant fatigue when presented with the game audio clips.

4.2.1 Audio Clip Selection. The audio clips contained samples of player interactions from the online first-person shooter, *Overwatch*. The three clips were amongst a collection of 50 videos sourced from Twitch streams, which were selected according to the following criteria: that the streamer is a woman; who uses voice chat in-game; is playing a competitive (ranked) game mode; and is in a party of three members or less. These criteria were established to strengthen the chances of encountering toxic social interactions when reviewing Twitch VODs (videos on demand): women, as they experience more harassment in games than men [27]; using voice chat, increasing social interaction; competitive game modes, where personal stakes are higher; and as a party of three members or less, ensuring the presence of at least three unknown teammates. The first author viewed Twitch VODs and recorded games that included instances of toxicity within voice chat. The first author then collected an initial sample of 50 clips, identified as containing potentially toxic interactions based on the following criteria: profanity and slurs directed towards others; bigoted comments and insults; malicious jokes or sarcastic remarks; players seeking to control others; and verbal aggression (e.g., arguing, yelling, or exhibiting frustration). The first author then selected 12 clips from this larger database, seeking disparity in toxicity based on the duration and severity of the interactions. Finally, the first and last author independently reviewed the 12 clips, and after conferral selected three clips representing low, medium, and high toxicity. To ensure the game communications were not influenced by the popularity or recognizability of the streamer, we selected streamers with a Twitch follower count that ranged from 1,600 to 37,300.

4.2.2 Audio Clip Summary. The three clips used were selected from the larger video pool for their comparatively short runtime (to minimize participant fatigue), disparity in toxic content, and legibility. We chose clips that the first and last author perceived as low (LowTox), medium (MidTox), and high (HighTox) in their level of toxicity. The clips had a runtime of 8:10 (LowTox), 7:25 (MidTox), and 7:45 (HighTox) minutes respectively, with non-relevant content (e.g., interactions between the Twitch streamer and their chat) muted to preserve clarity. The clips were limited to audio to remove the potential of additional variables (e.g., facial expressions) influencing perception of the game behaviors. We chose to exclude video content as audio is also largely how players would perceive the interactions (sans others' gameplay perspectives) in their own gameplay experience. Further, we wished to restrict perception of toxicity to the verbal interactions (uninfluenced by gameplay), and ensure focus remained on these.

LowTox: The behaviors within LowTox were limited to discussion of gameplay, such as strategizing (e.g., which characters to play) and callouts (e.g., alerting the team to an enemy's presence). Some discussion of strategy was terse, e.g., "I need you to not play Zarya on this map", but positive feedback was also provided (e.g., "Good job" after a player secures a kill). At one point, a player states that they have done the most damage in the team; a teammate sarcastically replies with, "You're good, you're good, you're really good". Communications after this exchange continue as before.

MidTox: Communications are initially amicable; one player greets the team, and interactions concern discussion of strategy and callouts. As the team starts to lose, players begin criticizing one another's gameplay and performance; one player states, "We could probably go *not* Widow on this one", and another responds by passive-aggressively suggesting the player uses their hero's abilities. As the match continues, comments devolve into attacks directed at performance, e.g., "Where'd you buy your account from, homie?" (implying that the player did not earn their high rank), "Struggling on the easiest role in the game, huh?", and, "If you can't get a single kill, what's the point?". Standard callouts and occasional positive feedback continue throughout. At the conclusion of the match, two

players tell each other that they will "avoid" (filter each other from matchmaking) one another.

HighTox: Communications are immediately hostile. One player jokingly flirts with another player, and is rebuked by a third and fourth player with, "That is really cringe", and "Dude, I'm glad you fucking said it so I didn't have to". The first player responds with, "Imagine not getting an obvious joke." The players temporarily focus on callouts and strategy, but devolve into personal and performance-directed attacks after some losses. Players use ableist slurs ("You play like a fucking retard"), gendered harassment ("You're a woman, didn't ask, didn't care"), and accuse one another of being bad at the game ("3.5 support player and you think you're hot shit"). There is consistent discussion of sexual themes; one player insinuates another does not have sexual experience, and should not defend a female player because, "She's not going to sleep with you, bro". The female player retorts by saying that she will, adding, "Maybe we'll send you the video of it, because clearly you're sexually frustrated". Players continue to provoke one another until the match ends.

4.3 Measures

Moral Disengagement in Games (MDG): We adapted a scale on Moral Disengagements in Sports [6] to competitive digital gaming. Originally comprised of six subscales: Conduct Reconstrual (CR; 8 items), Advantageous Comparison (AC; 4 items), Nonresponsibility (NR; 8 items), Distortion of Consequences (DC; 4 items), Dehumanization (DH; 4 items), and Attribution of Blame (AB; 4 items), two subscales (AC and NR) were removed as they were less relevant in the context of digital games. Each subscale is theoretically related to Bandura's [4] eight mechanisms of moral disengagement. Boardley and Kavussanu [6] created the CR subscale by merging the moral justification and euphemistic labelling mechanisms of moral disengagement as these factors both cognitively reconstrue behaviors as less harmful. The NR subscale was formed by combining the diffusion of responsibility and displacement of responsibility factors as both act by minimizing personal responsibility. The four remaining subscales (i.e., AC, DC, DH, and AB) are mechanisms of Bandura's [4] moral disengagement (see Section 2.3 for definitions). Included subscales, totalling 20 items, underwent minor semantic revisions (e.g., 'fouling' to 'flaming'). Statements (e.g., "Flaming an opponent is okay if it discourages them from flaming your teammates") were rated on a 7-point Likert scale. Results were averaged for each subscale with higher scores reflecting a greater tendency for moral disengagement in games. The scale had acceptable reliability: CR (α = .86), DC (α = .86), DH (α = .81), and AB (α = .73).

Online Disinhibition Scale: The 11-item ODS [55] was deployed to measure participant tendency towards online disinhibition along two subscales: Benign Disinhibition (7 items) and Toxic Disinhibition (4 items). The sample item "The Internet is anonymous so it is easier for me to express my true feelings or thoughts" reflects Benign Disinhibition, whereas "I don't mind writing insulting things about others online, because it's anonymous" reflects Toxic Disinhibition. Online disinhibition was assessed using a 4-point Likert scale. The reliability of the measures revealed acceptable Cronbach's alpha coefficients for each measure: Benign Disinhibition (α = .81) and Toxic Disinhibition (α = .85).

The Aggression Questionnaire: The 29-item Aggression Questionnaire [9] assesses tendency towards aggression. The measure is composed of four subscales: Physical Aggression (PA; 9 items), Verbal Aggression (VA; 5 items), Anger (A; 7 items), and Hostility (H; 8 items). Sample items from the subscales include: "Once in a while I can't control the urge to strike another person" from PA, and "I can't help getting into arguments when people disagree with me" from VA. The questionnaire was assessed on a 5-point Likert scale. The measures have adequate reliability for scales with fewer than 10 items, detailed as followed: PA (α = .85), VA (α = .72), A (α = .83), H (α = .77), and overall aggression (α = .89).

Perceived Toxicity: We used an 8-item scale from earlier research on social play in games [14]. On 7-point scales, participants were asked to rate their agreement whether they considered the people in the audio clip had different characteristics (e.g., *offensive*, *toxic*, and *angry*). They were instructed to consider the overall atmosphere of the social interactions in the audio clip. A single toxicity score was calculated from the individual items. Reliabilities were good to excellent across the three clips (*LowTox* α = .857, *MidTox* α = .918, *HighTox* α = .933).

Report Prompts: Two custom items were developed to assess both participant willingness to hypothetically report the behavior in the clip, and to evaluate why they would or would not report. Willingness to report was assessed with the item, "If you were playing in this match, would you report the behavior you heard in the clip?" (Yes, No, Unsure). Evaluation of their reporting decision was assessed with the pursuant open-ended item, "Why or why not?".

4.4 Participants

Data were cleaned to remove bots, incomplete attempts, and non-diligent respondents. Response variances were validated and responses to the toxicity prompts were analyzed to detect copy-pasted text fragments and obvious cases of non-diligent responses. With this step, we excluded data from 93 respondents.

The final sample consisted of data from 106 participants (men=68, women=35, non-binary=2, prefer not to disclose=1) aged 16 to 69 (M=32.7, Mdn=30.0, SD=9.3). All reported playing games, with the majority playing every day (N=54) or a few times per week (N=44). Participants answered a scale (0 - 100) that asked them to self-identify as a gamer; this scale has been validated against a 60-item questionnaire on self assessment of personal attributes in the domain of video game ability (r=.735). Our participants generally identified as gamers (mean = 68.9, SD = 30.4, min = 0, max = 100). As our study featured clips from *Overwatch*, participants completed a similar scale for *Overwatch* knowledge (mean = 47.0, SD = 28.4, min = 0, max = 100) and *Overwatch* skill (mean = 35.9, SD = 28.5, min = 0, max = 99), demonstrating that our participants were, in general, familiar with *Overwatch* (refer to Fig. 1 for the distributions of this data). Finally, participants were asked to report their peak rank in *Overwatch*: 46 had not obtained a rank, three had very high ranks (Masters = 1, Grandmasters = 2), and the remainder reported different ranks to similar degrees (in ascending order of rank: Bronze = 17, Silver = 10, Gold = 11, Platinum = 10, Diamond = 9), suggesting our dataset covered a satisfactory set of varied previous knowledge and skill.

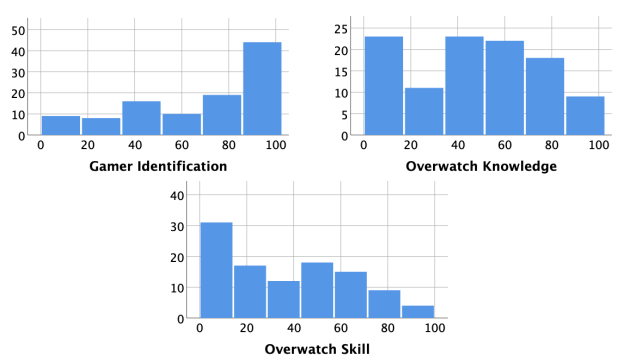

Figure 1: Distributions of participant gamer identification, Overwatch knowledge, and Overwatch skill.

5 RESULTS: PERCEPTIONS OF CLIP TOXICITY

First, we explored perceptions of toxicity across the three clips (see Fig.2 for distributions of ratings). Perceptions of toxicity differed as expected with the three clips being perceived as low (M=2.7), moderate (M=3.6), and high (M=4.8) in toxicity. A repeated-measures ANOVA confirmed this effect to be significant ($F(2, 210) = 93.731$, $p < .001$, $\eta_p^2 = .472$, Greenhouse-Geisser corrected). All pairwise comparisons using Bonferroni corrected post-hoc tests were highly significant (p <.001). The participants' willingness to report the behavior showed similar patterns. (LowTox: Yes=1, No=105; MidTox: Yes=9, Unsure=7, No=90; HighTox: Yes=45, Unsure=5, No=56). This in itself is interesting because it confirms that reporting options in games are useful to identify when toxicity occurs. We selected the HighTox clip to explore our research questions around perceptions of toxicity, and used the ratings of toxicity, as it is a continuous measure. A histogram of the toxicity ratings based on willingness to report for the HighTox clip (see Fig. 2) confirms that constructs are related, but that the toxicity ratings provide more granular insights into perceptions of toxicity.

Relationship Between Demographics and Perceptions of Toxicity. We calculated correlations between toxicity ratings and demographic factors and *Overwatch* background. Toxicity ratings were negatively, but not significantly correlated with age (r = -.179, p = .066) and gender (r = -.046, p = .639), but positively and significantly correlated with gamer identity (r = .242, p = .012) and *Overwatch* knowledge (r = .261, p = .006). In other words, people tend to rate the clip more toxic when they identified more strongly as gamers and had higher existing knowledge about *Overwatch*. To distill the relationship of toxicity ratings and traits, we controlled for those four variables.

The Relationship Between Traits and Perceptions of Toxicity. We conducted linear regression analyses using traits to predict toxicity ratings. Using hierarchical regressions, we controlled for age, gender, *Overwatch* knowledge, and gamer identification by entering them in the first block and adding predictors of interest in the second block. To avoid multicollinearity issues between individual predictors, we calculated separate regression models for each predictor. To control for multiple tests, we adjusted the significance

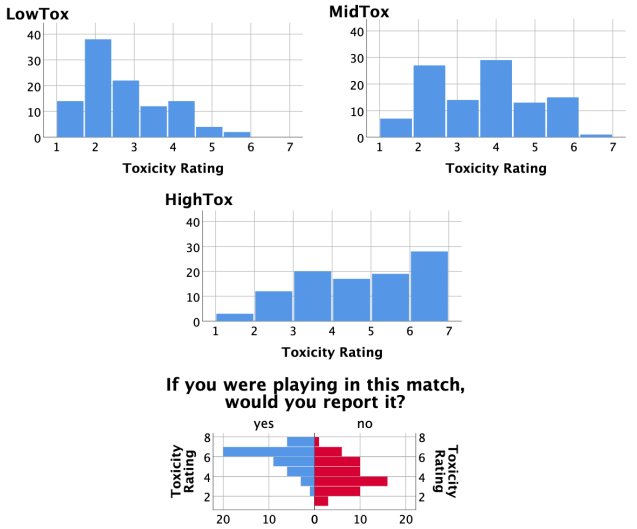

Figure 2: Distributions of toxicity ratings for each of the clips, and specifically split by whether or not players would report the behaviour for HighTox.

		B	p	R^2
MDG	Conduct Reconstrual	−0.376	.001**	.190
	Distorting Consequences	−0.330	.002**	.186
	Dehumanization	−0.486	<.001***	.251
	Attribution of Blame	−0.186	.098	.127
ODS	Benign Disinhibition	−0.106	.708	.104
	Toxic Disinhibition	−0.831	.001**	.198
AGG	Physical Aggression	−0.141	.475	.108
	Verbal Aggression	−0.156	.364	.110
	Anger	−0.283	.120	.125
	Hostility	−0.304	.090	.129

Table 1: Regression results for the prediction of toxicity ratings using moral disengagment (*MDG*), online disinhibition (*ODS*), and aggression (*AGG*) traits with unstandardized regression coefficients (B), p values, and explained variance (R^2). $^{**}p < .01$, $^{***}p < .001$

threshold using Bonferroni correction (divided by 10 for the number of regressions) and accepted effects as significant at $p < 0.005$.

The results (see Table 1) suggest toxic ratings had negative relationships with all traits. However, only four were significant predictors in the moral disengagement factors *conduct reconstrual*, *distorting consequences*, and *dehumanization*, as well as *toxic online disinhibition*. The other traits were not significant, considering the adjusted p-value threshold.

6 RESULTS: REPORTING THEMES

We conducted an inductive thematic analysis following Braun and Clarke [7]. Our study prompted participants to state their willingness to report behaviors in the clips using in-game report functionality (*"If you were playing in this match, would you report the behaviour you heard in this clip?"*), and to justify their reporting

decision (*"Why or why not?"*). All items were immediately coupled with the relevant clips to preserve participant recall.

The objective of the thematic analysis was to identify themes in the perception of toxicity in online gaming interactions, explored along the division of willingness to report. The last author reviewed all participant responses for the HighTox clip and iteratively conducted an initial coding of the data. Responses were categorized by willingness to report ('Yes', 'No', and 'Unsure'), and both latent (i.e., interpreting the participants' intended meaning) and semantic (i.e., the participants' words, verbatim) codes were generated. Responses could be assigned multiple codes. After the first coding phase, an initial set of codes and themes were generated, discussed amongst the authors, and regrouped and refined into a final set of themes. The final themes were not exclusive; codes could contribute to multiple themes simultaneously. We coded the full set of data to the established themes using a constant comparative analysis approach [7].

6.1 Themes: Contextualizing Reporting

'Contextualizing Reporting' explores the reasonings provided by participants who indicated their willingness to report the HighTox clip content. While the prompt did not ask participants to provide their long-term motivations for reporting, some divulged that they would be galvanized to report the behavior out of a need to protect their own, or others', experiences. In this analysis, we generated five recurring themes contextualizing participant decision to report. These themes are detailed, and supported by participant quotes (original spelling and grammar intact), in the following section.

Insults and Profanity. Participants indicated willingness to report HighTox content owing to the presence of verbal attacks on teammates, consisting of both personal- and performance-directed insults. Participants stated that the *"insults were a bit too extreme"* (P77), that *"they were all being so rude to each other, hurling around insults about skill levels, gender, etc."* (P14) and that the interactions were *"really mean. it was some really personal attacks"* (P31). Language also dictated willingness to report, with some participants citing *"bad words!"* (P28) and *"a lot of cursing words and dirty language"* (P39) as their rationale for reporting the content. The context of the interactions elevated the perceived severity or impropriety of the interactions for some; participants stated that, *"what they were talking about had nothing to do with the game"* (P52), and, *"I play games to enjoy myself and when someone goes out of their way to make it an unenjoyable experience it goes against what the game is for. Its not a place to bad mouth people or make sexual remarks"* (P44). As such, the insults and use of hostile language—elevated by the inappropriate context—motivated participants to report.

Hate Speech. Participants stated that they would report content that they perceived to be bigoted in nature; e.g., *"they made repeated sexist comments"* (P1), and *"And then we get into light ableism, which is… not desirable."*. This involves the use of ableist or sexist slurs (*"A lot of slurs against women being used, a lot of personal attacks, a lot of ableist slurs being used."* [P78]), which were regarded as offensive, ban-worthy, and unacceptable: *"Because of the use of 'fucking retard' it is offensive and intolerable"* (P59), *"Dude kept saying retard and that is reportable and should be banned"* (P42), *"there were a lot of rude comments about women, sex and using slurs."* For some participants, their own experiences in online gaming informed their decision to report. P76 stated that, *"as a female gamer, I'm so sick of people implying my friends are friends with me because I'm a woman"*; likewise, P73 explained that they would specifically report the player exhibiting sexist attitudes: *"Yeah the 'Well you're just a woman' dude is a jackass. She was crude but that's what women in gaming need to do, so the sexists and racists aren't allowed to fester."* While related to Insults and Profanity, the use of sexist and ableist slurs, stereotypes, and gender-driven harassment were noted extensively and disparately from general insults and cursing; it may be that, for some, bigotry represent a stronger impetus for reporting than non-bigoted aggression.

Bullying. Disparate from *Insults and Profanity* was the concept of bullying: targeted harassment of a single player—*"throwing around insults and picking on one player"* (P44). Participants highlighted that players in the clip targeted one particular player for not adhering to social expectations: *"They seemed to be bullying the guy for not understanding what everyone else claims was a joke ... it seemed to start off as kidding around but they didn't drop it calling the guy cringe and making fun of him"* (P1), *"right off the bat they were making fun of people and targeting certain people about not getting laid"* (P34). Further, some participants noted that players were *"targeting others based on their gender"* (P29) and that *"the female player is getting harassed"* (P6), with male players *"flaming"* the only female player *"for just being a girl"* (P49). For some participants, targeted abuse or harassment of a single player represented an impetus to report.

Sexual Content. Participants highlighted the presence of sexual content—satirical solicitation, explicit jokes, and descriptions of sex acts—as *"too aggressive and inappropriate"* (P30), and potentially harmful to other players: *"There was also some sexual pressuring going on, which could make a lot of people uncomfortable"* (P62). The discussion of sexual content in itself warranted reporting, with participants citing *"sexually explicit conduct"* (P1), *"the sexually explicit jokes"* (P6), and *"too much ... sexual talk"* (P67) as reasons to report. As with *Insults and Profanity*, the context of the interactions was important and elevated the inappropriateness of the interactions—to this end, P1 explained that the clip's content was *"somewhat vulgar and sexual and not the kind of discussion I want to listen to and quite annoying, and thus disrespectful to other players"*, with P102 stating that *"the talk should be focused on the game about winning."* In fact, the presence of sexual content was considered harmful to the experience—*"Too much sexual content for gamers. Detracts from the game."* (P46)—and represented an impetus to *"try and remove myself from it"* (P79).

Non-specific toxicity. This theme encapsulates allusions to non-specific toxicity, aggression, hurtful comments, and hostility. Participants alluded to *"alot of negative"* (P106) interactions that were *"a little over the top with the direction things took verbally"* (P106) and "*hurtful*" (P29). Many participants referred to general 'toxicity' and negative discussion as reason enough to report, e.g., *"it was highly toxic"* (P102), *"All players were being toxic"* (P47), *"the girl was complaining and talking a lot of crap"* (P49), and *"Just overall toxic behaviour from them"* (P88); this was potentially because these interactions contributed to *"a very negative and toxic atmosphere"*

(P3). Overall, non-specific toxicity, hostility, and the cultivation of an unpleasant atmosphere contextualized the decision to report for some participants.

6.2 Themes: Rationalizing Not Reporting

'Rationalizing Not Reporting' explores the rationales provided by participants who indicated that they would not report HighTox clip content. In this analysis, we generated four recurring themes—detailed, and supported by example codes, in the following section.

Banter. Prevalent in descriptions of the game interactions was the concept of 'banter': an acceptable form of trash talk considered to be jovial or humorous in nature, benign, and largely absent of malicious intent. To this end, the interactions were designated as *"playful banter"* (P97), *"messing around"* (P33), and *"just gamers talking smack"* (P36). Comparison was also key here: participants would allude to the playful or casual nature of the interactions, while likewise stating that, *"It's just swearing and some sexist talk. I've heard a LOT worse. The girls probably like them."* (P22), and *"mostly just some rudeness and some banter here and there not the worst thing"* (P51). These interactions—when characterized as banter—were in fact often perceived as beneficial to the players' experience, or as 'fun', e.g.: *"they are just having fun joking around"* (P98), *"I don't think reporting people for trash talking is worth it. Its part of what makes the game fun."* (P91), and *"there was nothing to be taken seriously outside of having fun"* (P66). This highlights that interactions that can be perceived as toxic or insulting by some (Do Report) may likewise be interpreted as playful, positive, and integral to a fun player experience by others (Don't Report).

Typical of Games. If the behavior is perceived as toxic, it's perceived as an inalienable aspect of communication within the parameters of online game interactions—and pointless to report. P15 states that, *"I wouldn't report this clip because this is not something new in online gaming and its not something that will be gone anytime soon. This has been the case since online lobbies even as far back as the first CoD Modern Warfare multiplayer."*, highlighting a belief that toxic interactions are an engrained element of online gameplay—that it's *"pretty typical for people to say very hard things when playing competitive games online"* (P87), *"not so uncommon"* (P100), and that *"you get used to it playing multiplayer games"* (P6). While the participants in this theme did not necessarily sanction the behavior, as they did in *Banter*, the context of online gameplay either normalized the interactions or rendered the action of reporting ineffectual.

Acceptable Toxicity. The behavior is perceived as toxic, but within acceptable thresholds or does not violate intangible or tangible rules (e.g., no one voices personal hurt, no threats of violence, no cheating). In these instances, participants noted that there wasn't *"anything that was extremely offensive"* (P101), or that the content was *"offensive, but not that bad"* (P24), with *"nothing severe enough to report"* (P50). P54 describes a threshold for reporting: *"unless someone did something to truly anger me, you know, being really aggressively nonstop toxic to other players, then I would report"*, and that the interactions in HighTox did not meet that threshold—*"while there is a lot objectionable here, nothing that really would have me itching to report"*. Perception of emotional harm also influenced willingness to report. P65 states that, while it was *"distasteful for me"*, they would not report as *"other people didn't seem to mind"*; likewise, participants noted that *"No one in the clip seemed to have hurt feelings or expressed they were upset"* (P20), and that *"if the friend group is fine with it, I have no problem with it"* (P68). Finally, participants also expressed a reticence to report unless in extreme circumstances; e.g., if *"someone's life was threatened"* (P63), if *"they had threatened harm to someone"* (P45), or if a player had *"doxxed anyone"* (P6) (in which 'doxxing' refers to publicly disseminating private or identifying information). Some likewise reserved reporting for explicit rule-breaking, e.g., if players were *"using cheating software"* (P45) or *"broke any TOS"* (P6). Overall, while toxicity in the clip's interactions was acknowledged, it was either perceived as not severe enough to report or as acceptable within the HighTox players' social dynamics.

Not my Circus. Not my circus, not my monkeys—the behavior is potentially perceived as toxic, but participants did not feel an obligation to report due to a lack of personal involvement or general disinterest in reporting. In this theme, the participants' direct involvement mandated their willingness to report; e.g., *"I wouldn't bother reporting something like that, especially since I wasn't the one being addressed anyway"* (P2), or were reluctant to involve themselves, e.g., *"It doesn't seem like any good could come of trying to interfere in the relationships this group have"* (P101), *"I'll let them have their fun"* (P56). Others noted that they *"wouldn't care enough to report it"* (P11), with P58 stating, *"whooooooo cares about video game losers"*. Furthermore, some participants indicated that using the in-game report functionality in general was *"just not my style"* (P94), and that they have *"never reported someone for what they've said. I would just mute them."* (P19). Overall, this theme points to a lack of personal involvement or interest, or personal philosophy towards using in-game report functionality, mandating a disinclination to report.

7 RESULTS: TRAITS AND RATIONALIZATION

To explore who is most at risk of normalizing toxic behaviors in games, we categorized the rationalization themes by the traits that influenced perceived toxicity, using median splits. We opted to explore moral disengagement and toxic online disinhibition, but not aggression, as these contained traits that had significant relationships with perception of clip toxicity. As such, we suspected that moral disengagement and toxic online disinhibition may inform participant rationale for not reporting.

The qualitative themes were volunteered and inductive (i.e., participants were not asked about their opinion on each theme), thus participants may have agreed that it was just 'banter' had they been asked explicitly, but did not volunteer that rationalization in their free-text response. Consequently, we do not conduct statistical tests on these data, but interpret the patterns descriptively.

Fig. 3 shows the count of participants who reported each rationalization theme, split in terms of low and high moral disengagement (overall) and toxic online disinhibition. We included all the subscales of moral disengagement together as they were all negatively associated with perceived toxicity, and we did not include benign online disinhibition as it showed no significant association with perceptions of toxicity. As can be seen, although both traits previously predicted reduced perceptions of toxicity, when we consider

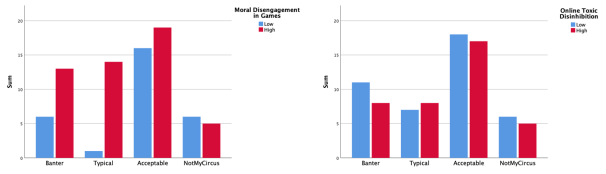

Figure 3: Count of participants who reported each theme, using median splits of MDG (Left) and online toxic disinhibition (Right).

the rationalization themes, the justifications 'Typical of Games' and 'Banter', appear to depend on moral disengagement, whereas there is no obvious relationship with online disinhibition. These results provide further support that when addressing normalizing of toxic behaviors in games, moral disengagement is a valuable trait to consider.

8 DISCUSSION

8.1 Overview

Overall, our findings demonstrate that participants who report a higher tendency towards Conduct Reconstrual (MDG), Distorting Consequences (MDG), Dehumanization (MDG), and Toxic Online Disinhibition (ODS) were less likely to perceive online game interactions as toxic. We found no significant relationships for Attribution of Blame (MDG), Benign Online Disinhibition (ODS), and Aggression. A thematic analysis of reporting justifications for HighTox revealed that participants who do report behavior point to toxic player behaviors (i.e., Insults and Profanity, Hate Speech, Bullying, Sexual Content, and Non-Specific Toxicity), whereas participants who don't report rationalize these same behaviors as contextually appropriate (Typical of Games), acceptable (Acceptable Toxicity), enjoyable (Banter), or not their problem (Not My Circus). Additional analysis also suggested that participants who rated higher for MDG traits were more likely to rationalize negative behaviors according to the themes Banter, Acceptable Toxicity, and Typical of Games.

8.2 How Traits Predict Perception of Toxicity

To investigate RQ1 (*"Are there traits that predict how toxic an interaction is perceived to be?"*), we explored predictive relationships between Moral Disengagement in Games, Online Disinhibition, and Aggression. Our findings reveal a negative relationship between Conduct Reconstrual (MDG), Distorting Consequences (MDG), Dehumanization (MDG), and Toxic Online Disinhibition (ODS), and the perceived toxicity of the HighTox clip. As such, participants with a greater tendency towards these traits were less likely to perceive the behaviors as toxic. The influence of moral disengagement on the perception of toxicity is noteworthy: the construct was introduced to help explain behaviors that deviate from personal morals [4], but has not yet been applied to toxicity in online games. A closer investigation of these traits helps us to understand factors contributing to the perception, rationalization, and normalization of toxicity in online gaming.

Participants who reported a greater tendency towards Construct Reconstrual rated the behaviors within all three audio clips as less toxic. The Construct Reconstrual subscale concerns a cognitive reconstrual of negative behaviors into positive ones, and the euphemistic rebranding of negative labels as positive [6]. This reconstrual occurs because the behavior may be perceived as having a valuable social purpose. In this case, players may morally justify hostile or abusive communications as beneficial, socially acceptable, or appropriate. The behaviors are further reframed and justified by adopting language that reduces the severity of the actions, such as "banter" when referring to insults or denigration. This aligns neatly with the themes generated for 'Rationalizing Not Reporting': players minimize behaviors as 'Typical of Games' (the context makes it appropriate) and 'Banter' (reframing behavior as beneficial, positive, or fun; what is perceived by some as insults and bullying, is instead interpreted as *"banter"*, *"trash talk"*, and *"joking around"*). Reframing behaviors to serve a moral purpose may serve to normalize them as acceptable within the context of online gaming, reducing the perception of toxicity. As the theory of normalized behavior suggests that those who are approving are more likely to engage in that behavior [24], it may be that Construct Reconstrual—and tacit approval of behaviors interpreted as 'banter'—may be a factor behind the cyclical perpetuation of toxicity in online games.

A greater tendency towards Dehumanization was also mapped with decreased perceptions of toxicity. Dehumanization concerns a tendency to deprive others of their humanity, resulting in reduced empathy or visualization of the potential for emotional harm [6]. In online gaming, this effect could potentially be magnified by the online disinhibition effect (other players are anonymous, distanced, and two-dimensional) [51]. As such, negative behaviors may be perceived as less egregious owing to the dehumanization of other players; this may especially occur in instances in which players are reduced to a collection of negative traits and actions (e.g., poor performance or hostile communication). Following this, treating these players inhumanely is considered justified—decreasing perception of toxicity. Dehumanization of the players may have also been amplified by the study design: the participants weren't players in a real-time scenario, but were instead listening to pre-recorded audio of others. This extra level of obfuscation may have contributed to further dehumanization and normalization of toxicity.

The final significant finding for MDG was an increased tendency towards Distorting Consequences predicting decreased perception of clip toxicity. Distorting Consequences occurs when the individual cognitively minimizes the harm caused by their action, but can be subverted when the recipient demonstrates hurt or suffering [4]. By minimizing the harm caused to others, their perceptions of toxicity are also minimized, as they do not view their behaviors as harmful. This maps with the 'Acceptable Toxicity' theme generated, in which participants state that the negative behaviors were acceptable because the players did not seem (or explicitly state that they were) upset or otherwise harmed by the negative behaviors. This theme also introduces an additional theoretical threshold for when the participants might view the clip as toxic: participants claim that they would not be concerned unless someone was threatened or compromised—explicit examples of harm that are potentially less susceptible to minimization. This suggests denigrating or negative behaviours may be normalized as 'acceptable'—until, or unless,

recipients explicitly verbalise that they are hurt by the toxic behaviors.

In terms of the ODS, we found that Toxic Online Disinhibition was a meaningful predictor of decreased toxicity perception. This supports (and confirms in a different genre) findings from recent literature, which finds that online disinhibition contributes to toxic behaviors in MOBAs [26]. These findings suggest that those who are more disinhibited towards toxic behaviors online—because they are anonymous, disassociated, or distanced from the consequences of their actions—will likewise perceive potentially negative behaviors performed by others, unto others, as less toxic.

We did not find any meaningful results for Benign Online Disinhibition (ODS), Attribution of Blame (MDG), and Aggression. While the distinction between toxic and benign online disinhibition is ambiguous [51], benign disinhibition manifests as openness, kindness, and generosity; however, the subscale does not concern attitudes towards negative online behaviors. It may be that individuals who trend towards benign disinhibition are equally capable of perceiving toxicity as more (it is not kind, open, or generous) or as less severe, because of the online context. Attribution of Blame occurs when people see themselves as victims provoked into negative action; however, just because someone else may have "started it" does not necessarily imply that the players (or participants) perceive retaliatory behaviors as less toxic. Additionally, the participants in this study were not playing the game (nor were they the target of toxicity)—people may self-victimize to defend an action, but a third party may be less likely to afford this justification for others. Finally, while research suggests that aggression is normalized within online gaming communities [23], that does not imply that the players who normalize it possess aggressive tendencies. Continuous exposure to aggression in online gaming communities, rather than aggressive tendencies themselves, may provide a better explanation for the normalization of antisocial behaviors in the context of video games.

8.3 Contrasting Reporting and Not Reporting Themes

That would/wouldn't report for HighTox was nearly evenly divided (Yes = 45, No = 56) suggested different perceptions of, or attitudes towards, the same content. To address RQs 2 and 3 (*"What informs participants' decisions to report toxic behaviors?"*, and *"How do players rationalize not reporting observed toxic behaviors?"*), we explored the justifications participants provided for reporting or not reporting toxic behaviors. The themes for 'would report' and 'wouldn't report' reveal two disparate approaches towards informing participant decision: while participants who *do* report generally highlight discrete or specific behaviors as their impetus (e.g., bullying, sexual content), participant who *don't* report instead largely rationalize the behaviors (e.g., banter, typical of games) or absolve themselves of hypothetical responsibility (not my circus).

The themes generated for *'Contexualizing Reporting'* supports what is known about toxicity: that insults and profanity, hate speech, targeted bullying, sexual content, and non-specific toxicity (e.g., negativity, a toxic atmosphere) are forms of negative or toxic behaviors [19, 21, 29, 54]. That some participants would identify these behaviors as toxic, and be compelled to report them, aligns with the literature. Interestingly, the context of an online gaming space actually informed participants' impetus to report. In *Insults and Profanity*, participants state that the toxicity *"goes against what the game is for"*, and that it has *"nothing to do with the game"*. This sentiment was echoed in *Sexual Content*.

In contrast, the majority of the themes generated for *'Rationalizing Not Reporting'* suggested that the context of online gaming instead rendered the behaviors positive, appropriate, or tolerable—normalizing them. *Banter* reimagines abusive or negative interactions as playful banter, ribbing, or joking; acceptable in competitive environments, and not to be taken seriously. *Typical of Games* likewise suggests that such interactions are an inalienable and intrinsic element of online gaming; that toxicity is typical or standard, and something to get used to. *Acceptable Toxicity* suggested that the behavior was relatively mild for online games, and while potentially objectionable, did not represent something worth reporting. Taken together, these three themes suggest the normalization of toxic behaviors in online games.

Both *Acceptable Toxicity* and *Banter* speak to the potential insidiousness of negative online behaviors. As our participants did not observe overtly aggressive interactions (e.g., threatening harm or revealing sensitive information), they did not perceive the behaviors as toxic or feel compelled to report them. As many instantiations of toxicity are more covert or underhanded than this (e.g., insults disguised as banter), this may contribute to the normalization of toxicity in games, and explain its pervasiveness. This may also be amplified by trait tendency towards Distorting Consequences: unless the consequences of the negative behavior are made explicit, the negative behavior is rendered as less harmful.

Interestingly, the fourth theme generated for *'Rationalizing Not Reporting'*, *Not My Circus*, does not seem to suggest normalization of the behaviors—but rather, rationalization of the participant's non-response. In this theme, participants describe a hypothetical hesitation to involve themselves (*"I wasn't the one being addressed anyway"*), or describe a general unwillingness to use report functionalities (*"not my style"*). This theme suggests that, while participants do recognize the behavior as potentially toxic or inappropriate, a lack of personal involvement or general philosophy to reporting stymies their willingness to report. This suggests that uninvolved players—those not directly affected by the hostility—are less compelled to take action against it. A bystander effect (in which a player's sense of responsibility to report is diffused by the presence of others) may also inform reticence to report. The presence of this effect has been theorized as a contributing factor towards unwillingness to report in previous toxicity research [2, 5, 28].

8.4 Traits and Rationalizations

To explore RQ4 (*"How do traits relate to the rationalization of toxic behaviors in a gaming context?"*), additional exploratory analyses were performed that suggested that participants who rated higher for MDG traits were also more likely to rationalize negative behaviors according to the themes *Banter*, *Typical of Games*, and possibly (albeit less clearly) *Acceptable Toxicity*. While we should be cautious in the interpretation of this exploratory analysis, this does suggest a relationship between increased Moral Disengagement in Games and themes of normalization. This is also supported by parallels between the traits and the themes discussed in Section 8.2.

The distinction between *Not My Circus* was not as clear, with participants in the 'low' grouping for Moral Disengagement in Games reporting *Not My Circus* at a slightly higher rate. This fuzzy distinction is not surprising: unlike the other three themes generated for 'Rationalizing Not Reporting', *Not My Circus* did not possess motifs that were clearly related to normalization. Instead, the theme largely concerned personal non-responsibility and attitudes towards the report functionality in general; these are sentiments that could be shared by those who both trend towards, and away from, moral disengagement in games.

8.5 Summary

Taken together, these findings further support that toxicity in online gaming is a complex and multi-faceted problem. The disparity in perceptions of and attitudes towards identical content demonstrating negative behaviors further highlights the complexity of identifying, addressing, and eradicating toxicity in online games. This research demonstrates that the same behaviors that are perceived as inappropriate, intolerable, and harmful by some can be interpreted as appropriate, tolerable, and enjoyable by others. We find that player traits play a role in this, and that those who trend towards moral disengagement in games and toxic online disinhibition are more likely to rationalize negative behavior. We also find that normalization is endemic to the rationalization of toxic behaviors; toxicity is considered, by some, as an unavoidable and acceptable feature of online gaming—and that those who trend towards moral disengagement in games are more susceptible to normalizing toxic behaviors. As normalization of a behavior begets that behavior, this presents a critical issue and an entry point for interrupting cyclical toxicity. We also propose that non-responsibility, and unwillingness to involve oneself, and personal reticence towards using the report functionality plays a role in reporting behaviors in online games. We further suggest, in concert with extant literature [2, 5], that some reluctance towards reporting may be cautiously attributable to the bystander effect.

This research helps inform our understanding of what constitutes toxicity, how players perceive it, and the mechanisms behind normalization. It is crucial for researchers and developers to understand not only what players want, but who they are; this is critical for understanding community attitudes towards toxicity and informing design decisions to discourage negative behaviors but encourage reporting them. By recognizing the importance of different traits, design elements can be implemented to combat the normalization of toxicity. We also suggest that normalization is an embedded cultural issue within online gaming, and that normalization emerges both out of tacit approval of negative behaviors and resignation towards them; ergo, developers should remain cognizant of their game community's attitudes towards toxicity when implementing frameworks that detect, punish, and sanitize (e.g., censor or filter) negative behaviors.

8.6 Implications

Our research highlights that the normalization of toxic behaviors is an embedded cultural issue that may be reinforced by player traits. By recognizing how different traits interact with online environments such as video games, design elements can be implemented to combat the normalization of toxicity—especially for those susceptible to perceiving toxicity as acceptable behaviors. For example, developers could inform players of behaviors that are considered toxic early on in the introduction of the game—during the tutorial, or first few hours of play—to break the cycle of toxicity and normalization. Alternatively, developers may wish to subvert Distorting Consequences by demonstrating the negative consequences of a player's actions (e.g., a recipient indicating that they were hurt).

We also suggest that unwillingness to report is a consequence of diffusion of responsibility (i.e., a bystander effect), not wanting to get involved, or personal disinterest in using in-game report functionality. This represents a systemic issue that warrants further exploration by developers and researchers—for example, how we might disrupt bystander effects, or incentivize reporting toxicity.

Our work informs possible concrete solutions that may benefit from consideration or implementation in competitive multiplayer contexts. We note that Sparrow et al. suggest that normalization may stem from a lack of faith in reporting systems [47]; to this end, we posit that cultivating open, transparent communication with playerbases (e.g., revealing what is and isn't working, and providing evidence of progress) can help to establish a relationship of trust. This may allow developers to rebuild player faith and agency in moderation infrastructure, and could prove especially motivating for players who opt not to report for reasons detailed in *Typical of Games*—wherein players alluded to a perceived pointlessness in reporting.

Further, it may be helpful to scrutinise player-described motivations *for* reporting. One such motivation for reporting was a desire to 'protect others' (e.g., young players, the community); following this, we suggest a system or tool that positions reporting and moderation as a critical community-driven protective effort with tangible results. One example of a prior implementation of this concept is the *League of Legends* Tribunal system, wherein players could review reported chat content and decide upon an appropriate punishment (if any) for the accused player. While this system was ultimately retired as it was "slow" and "inaccurate", it was nonetheless acknowledged as supportive of player agency in community moderation [38]. Thoughtful iteration upon a system similar to this—for example, improved integration with the game client, visible and community facing cosmetic incentives, or collaboration with employed moderators—may help to restore player faith in reporting systems, and energize player involvement in preserving a healthy online community. To this end, we suggest that this may be an beneficial system to incorporate regardless of its efficacy in actual community moderation.

9 LIMITATIONS AND FUTURE WORK

In our analysis, we investigated how players interacted in a single game (*Overwatch*) and how observers perceived those interactions that occurred in a single match. As such, our results might be limited to the specific characteristics of *Overwatch* and in part the particular match that we used. For example, the HighTox clip did not contain obvious bigotry concerning race-ethnicity—which has been highlighted as a common form of abuse in online games [20, 21]. Thus, the themes that we generated in the qualitative analysis are

by no means exhaustive, while the quantitative analysis uses cross-sectional data and cannot be used to infer causality. Therefore, while our findings provide novel and exploratory insights, further research is necessary to confirm these results in similar and different contexts.

While we chose to restrict the toxicity clips to audio content only, the use of visuals poses an interesting avenue of investigation. Although not the focus of our study, observing the streamer's facial expressions from video content may allow participants to perceive the emotional impact of the toxic interactions from a third-person perspective. While players would not typically be exposed to this information in play, future work examining the effects of facial expressions on perceptions of toxicity may provide additional insight into how players rationalize toxic behaviors.

We did not find any significant results for Aggression, which was studied in the context of overt acts of aggression (e.g., verbal and physical aggression, hostility, and anger). Yet, as acts of aggression online may be more passive-aggressive, covert, or insidious in nature, future work that utilizes a scale that addresses this distinction may be beneficial.

One interesting result was that those who identified more strongly as gamers—and who reported more *Overwatch* knowledge—rated the HighTox clip was more toxic. One possible explanation for this may be that these participants were more likely to recognise toxic interactions, due to previous personal experience; alternatively, they may be less likely to dismiss online toxic interactions as unimportant or frivolous owing to the gaming context. Future work that investigates gamer identity and detection, reception, and understanding of toxic behaviours represents a cogent avenue for further investigation.

Our study addressed a relatively controversial topic in online gaming communities. As with most research investigating negative or controversial traits and concepts, participants may have sought to provide 'socially desirable' participant responses. While this should not influence the interpretation of the results, it is beneficial to remain cognizant of this potential limitation. Additionally, participants had the ability to skip through audio clips; while we instructed participants to listen to the full length of the clip prior to answering, this was not enforceable. Despite this, there is no evidence that participants skipped clip content.

In the preceding section, we have proposed several solutions would benefit from additional scholarly exploration and consideration—or implementation in competitive multiplayer contexts. In particular, these solutions concern disrupting the normalization of toxic behaviors through asserting their negative effects, restoring player faith in reporting and moderating systems, and facilitating player agency through active involvement in community preservation. We contend that HCI scholars can support ongoing approaches to normalization disruption through investigation of efficacy, and the generation of strategies for their implementation.

We highlight that trait tendency towards moral disengagement in games is a potentially important contributing factor towards the normalization of toxicity. As this has not been previously explored within games user research, we adapted a scale utilized for the exploration of this concept in sports. Based on our results, we propose that moral disengagement in games represents a critical avenue for continued investigation; and further, that it would be beneficial to formally develop (or adapt) a scale that allows us to continue to evaluate this construct within games.

10 CONCLUSION

In this paper, we investigated the perceptions of toxicity, reasons and rationalizations for reporting or not reporting toxicity in online games, and the role that player traits play in both. Consequently, our work contributes important new information concerning the normalization of toxicity in video games. We found that Moral Disengagement in Games and Toxic Online Disinhibition both contribute to the reduced perception of *how* toxic behaviors are. Furthermore, we identified normalization as a driving factor for the rationalization, tolerance, or acceptance of toxicity in online gaming—and find that players who trend towards Moral Disengagement in Games may be more susceptible to this normalization. Pursuant to this, we argue that the normalization of toxicity is endemic within online gaming spaces, is influenced by player traits, and represents a critical avenue for future work. We provide several recommendations for ongoing investigation, including continued research concerning normalization and moral disengagement in games. Research that builds upon this, combined with development efforts towards interrupting or dismantling the normalization of toxicity, will equip developers and researchers alike with the tools and knowledge to improve the health of online gaming spaces and communities.

ACKNOWLEDGMENTS

Thank you to our participants for their participation in the study, and to the University of Saskatchewan Interaction Lab for support.

REFERENCES

[1] Kate Abramson. 2014. Turning up the Lights on Gaslighting. *Philosophical Perspectives* 28, 1 (2014), 1–30. https://doi.org/10.1111/phpe.12046

[2] 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 (CHI PLAY '18 Extended Abstracts)*. Association for Computing Machinery, Melbourne, VIC, Australia, 365–372. https://doi.org/10.1145/3270316.3271545

[3] Mary Elizabeth Ballard and Kelly Marie Welch. 2017. Virtual warfare: Cyberbullying and cyber-victimization in MMOG play. *Games and Culture* 12, 5 (2017), 466–491.

[4] Albert Bandura. 1999. Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review* 3, 3 (Aug. 1999), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

[5] 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, Seoul, Korea, 877–888. https://doi.org/10.1145/2566486.2567987

[6] Ian D. Boardley and Maria Kavussanu. 2007. Development and validation of the Moral Disengagement in Sport Scale. *Journal of Sport & Exercise Psychology* 29, 5 (2007), 608–628. https://doi.org/10.1037/e548052012-028

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Brenda Jo Bredemeier and David L. Shields. 1986. Game Reasoning and Interactional Morality. *The Journal of Genetic Psychology* 147, 2 (June 1986), 257–275. https://doi.org/10.1080/00221325.1986.9914499

[9] Arnold H. Buss and Mark Perry. 1992. The Aggression Questionnaire. *Journal of Personality and Social Psychology* 63, 3 (1992), 452–459. https://doi.org/10.1037/0022-3514.63.3.452

[10] Marcus Carter and Fraser Allison. 2019. Guilt in DayZ. In *Transgression in Games and Play*, K. Jorgensen & F. Karlsen (Ed.). MIT Press, Cambridge, MA, Chapter 8, 134–152.

[11] Vivian Hsueh-Hua Chen, Henry Been-Lirn Duh, and Chiew Woon Ng. 2009. Players who play to make others cry: the influence of anonymity and immersion.

In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology (ACE '09)*. Association for Computing Machinery, Athens, Greece, 341–344. https://doi.org/10.1145/1690388.1690454

[12] Shira Chess and Adrienne Shaw. 2015. A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media* 59, 1 (2015), 208–220.

[13] Christine Cook, Juliette Schaafsma, and Marjolijn Antheunis. 2018. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society* 20, 9 (2018), 3323–3340. https://doi.org/10.1177/1461444817748578

[14] Ansgar E. Depping, Colby Johanson, and Regan L. Mandryk. 2018. Designing for Friendship: Modeling Properties of Play, In-Game Social Capital, and Psychological Well-Being. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 87–100. https://doi.org/10.1145/3242671.3242702

[15] Entertainment Software Association. 2019. Essential Facts About the Computer and Video Game Industry.

[16] S. Eveslage and K. Delaney. 1998. Talkin' trash at Hardwick high: a case study of insult talk on a boys' basketball team. *International Review for the Sociology of Sport* 33, 3 (1998), 239–253. https://www.cabdirect.org/cabdirect/abstract/19981811828

[17] Fair Play Alliance. 2020. Fair Play Alliance. https://fairplayalliance.org/

[18] Chek Yang Foo and Elina M. I. Koivisto. 2004. Defining grief play in MMORPGs: player and developer perceptions. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology (ACE '04)*. Association for Computing Machinery, Singapore, 245–250. https://doi.org/10.1145/1067343.1067375

[19] Jesse Fox and Wai Yen Tang. 2014. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Computers in Human Behavior* 33 (2014), 314–320.

[20] Kishonna L Gray. 2012. Deviant bodies, stigmatized identities, and racist acts: Examining the experiences of African-American gamers in Xbox Live. *New Review of Hypermedia and Multimedia* 18, 4 (2012), 261–276.

[21] Kishonna L Gray. 2014. *Race, gender, and deviance in Xbox live: Theoretical perspectives from the virtual margins*. Routledge, New York, NY, USA.

[22] Tilo Hartmann. 2017. The 'Moral Disengagement in Violent Videogames' Model. *Game Studies* 17, 2 (2017).

[23] Zorah Hilvert-Bruce and James T. Neill. 2020. I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior* 102 (Jan. 2020), 303–311. https://doi.org/10.1016/j.chb.2019.09.003

[24] L Rowell Huesmann and Leonard D Eron. 1984. Cognitive processes and the persistence of aggressive behavior. *Aggressive behavior* 10, 3 (1984), 243–251.

[25] Christoph Klimmt, Hannah Schmid, Andreas Nosper, Tilo Hartmann, and Peter Vorderer. 01 Sep. 2006. How players manage moral concerns to make video game violence enjoyable. *Communications* 31, 3 (01 Sep. 2006), 309 – 328. https://doi.org/10.1515/COMMUN.2006.020

[26] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* 30, 4 (April 2020), 1081–1102. https://doi.org/10.1108/INTR-08-2019-0343

[27] Jeffrey H. Kuznekoff and Lindsey M. Rose. 2013. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society* 15, 4 (June 2013), 541–556. https://doi.org/10.1177/1461444812458271

[28] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 3739–3748. https://doi.org/10.1145/2702123.2702529

[29] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28, 2 (March 2012), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

[30] Sung Je Lee, Eui Jun Jeong, and Joon Hyun Jeon. 2019. Disruptive behaviors in online games: Effects of moral positioning, competitive motivation, and aggression in "League of Legends". *Social Behavior and Personality: an international journal* 47, 2 (2019), 1–9.

[31] Regan L. Mandryk, Julian Frommel, Ashley Armstrong, and Daniel Johnson. 2020. How Passion for Playing World of Warcraft Predicts In-Game Social Capital, Loneliness, and Wellbeing. *Frontiers in Psychology* 11 (2020), 2165. https://doi.org/10.3389/fpsyg.2020.02165

[32] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (March 2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6

[33] Topias Mattinen and Joseph Macey. 2018. Online Abuse and Age in Dota 2. In *Proceedings of the 22nd International Academic Mindtrek Conference (Mindtrek '18)*. Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/3275116.3275149

[34] Newzoo. 2019. 2018 Global Games Market Report.

[35] Ryan Perry, Anders Drachen, Allison Kearney, Simone Kriglstein, Lennart E Nacke, Rafet Sifa, Guenter Wallner, and Daniel Johnson. 2018. Online-only friends, real-life friends or strangers? Differential associations with passion and social capital in video game play. *Computers in Human Behavior* 79 (2018), 202–210.

[36] Chrisa D. Pornari and Jane Wood. 2010. Peer and cyber aggression in secondary school students: the role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior* 36, 2 (2010), 81–94. https://doi.org/10.1002/ab.20336

[37] D.W. Rainey and V. Granito. 2010. Normative rules for trash talk among college athletes: an exploratory study. *Journal of Sport Behavior* 33, 3 (2010), 276–294.

[38] Riot. 2019. Ask Riot: Will Tribunal Return?

[39] L Rowell Huesmann. 1988. An information processing model for the development of aggression. *Aggressive behavior* 14, 1 (1988), 13–24.

[40] Anastasia Salter and Bridget Blodgett. 2012. Hypermasculinity & dickwolves: The contentious role of women in the new gaming public. *Journal of broadcasting & electronic media* 56, 3 (2012), 401–416.

[41] Daniel M Shafer. 2014. Moral Choice in Video Games: An Exploratory Study. *Media Psychology Review* 5, 1 (2014).

[42] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (July 2020), 1–6. https://doi.org/10.1016/j.chb.2020.106343

[43] David Light Shields, Christopher D. Funk, and Brenda Light Bredemeier. 2015. Predictors of Moral Disengagement in Sport. *Journal of Sport and Exercise Psychology* 37, 6 (Dec. 2015), 646–658. https://doi.org/10.1123/jsep.2015-0110

[44] Kenneth B. 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 (CSCW '14)*. Association for Computing Machinery, Baltimore, Maryland, USA, 1356–1365. https://doi.org/10.1145/2531602.2531724

[45] Catherine R. South and Jane Wood. 2006. Bullying in prisons: the importance of perceived social status, prisonization, and moral disengagement. *Aggressive Behavior* 32, 5 (2006), 490–501. https://doi.org/10.1002/ab.20149

[46] Spam. 2020. Merriam-Webster's Collegiate Dictionary. https://www.merriam-webster.com/dictionary/spam

[47] Lucy Sparrow, Martin Gibbs, and Michael Arnold. 2019. Apathetic Villagers and the Trolls Who Love Them: Player Amorality in Online Multiplayer Games. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Fremantle, WA, Australia) *(OZCHI'19)*. Association for Computing Machinery, New York, NY, USA, 447–451. https://doi.org/10.1145/3369457.3369514

[48] Nicholas Stanger, Maria Kavussanu, Ian D. Boardley, and Christopher Ring. 2013. The influence of moral disengagement and negative emotion on antisocial sport behavior. *Sport, Exercise, and Performance Psychology* 2, 2 (2013), 117–129. https://doi.org/10.1037/a0030585

[49] Cynthia A. Stark. 2019. Gaslighting, Misogyny, and Psychological Oppression. *The Monist* 102, 2 (April 2019), 221–235. https://doi.org/10.1093/monist/onz007

[50] Ashley E. Stirling and Gretchen A. Kerr. 2008. Defining and categorizing emotional abuse in sport. *European Journal of Sport Science* 8, 4 (July 2008), 173–181. https://doi.org/10.1080/17461390802086281

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

[52] Sabine Trepte, Leonard Reinecke, and Keno Juechems. 2012. The social side of gaming: How playing online computer games creates online and offline social support. *Computers in Human behavior* 28, 3 (2012), 832–839.

[53] Anna K. Turnage. 2007. Email Flaming Behaviors and Organizational Conflict. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 43–59. https://doi.org/10.1111/j.1083-6101.2007.00385.x

[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 (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376191

[55] Reinis Udris. 2014. Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior* 41 (Dec. 2014), 253–261. https://doi.org/10.1016/j.chb.2014.09.036

[56] Saulius Šukys and Aušra Janina Jansonienė. 2012. Relationship between Athletes' Values and Moral Disengagement in Sport, and Differences Across Gender, Level and Years of Involvement. *Baltic Journal of Sport and Health Sciences* 1, 84 (2012), 55–61. https://doi.org/10.33607/bjshs.v1i84.300