Combatting Toxicity: Designing an Intelligent System to Diminish Verbal Harassment in Online Games

by Jessica Patel

A thesis exhibition presented to OCAD University
in partial fulfillment of the requirements for the degree of
Master of Design in Digital Futures.

Exhibition: OCAD U Waterfront Campus, 130 Queens Quay E, March 27th - April 2nd
Toronto, Ontario, Canada, 2025

Copyright Notice

Abstract

The world of online multiplayer gaming has reached a point where player communities have become plagued with verbal toxicity in the form of hate and harassment, a detrimental issue for both players and game development companies alike. This master's thesis involves the research, design and development of an intervention system that can be adapted and integrated into online games. The goals of the intervention system are to detect, moderate, and potentially prevent verbal toxicity. Design thinking was used as a methodology to promote player-centered interfaces and interventions. Deep learning-based natural language processing (NLP) techniques were used to develop the back-end toxicity detection. The NLP algorithms process speech to analyze two modalities of verbal toxicity: text transcription and audio features. Investigation of audio feature analysis was prioritized since it is less prevalent in the existing literature and intervention systems. Audio feature extraction was explored to identify an optimal set of features that are both measurable and indicative of toxicity. The intervention system responds to detected toxicity by overlaying interface elements during gameplay to moderate and prevent verbal toxicity while accommodating diverse scenarios and player needs. User experience heuristics in learning, feedback, visual appearance and interaction were applied alongside player testing with 10 participants to develop a refined prototype that prioritizes strong player experience, player-system cooperation, and operant conditioning to correct toxic behavior.

Keywords: Verbal Toxicity; Hate and Harassment; Online Multiplayer Games; Moderation; Machine Learning; Natural Language Processing; User Experience; Human-Computer Interaction; Ethics

Contributions

This thesis was done with contributions from Yun Seok Yang, an undergraduate researcher in the Department of Engineering Science at the University of Toronto. I could not have done it without you!

This thesis was supported by the Ontario Graduate Scholarship, OCAD University, and the William and Phyllis Waters Bursary.

Acknowledgments & Dedication

My master's thesis has been one of the best journeys of my life. To feel so passionate, fulfilled, and excited about such an interdisciplinary project, one that makes me think so much, has been such a blessing. I truly appreciate each and every person who participated in the user research and critique sessions of my thesis. Puzzling over all the ethical dilemmas and unique challenges of toxicity truly opened my mind and drove me. This was truly one of my favorite parts of this experience.

I am incredibly grateful for all the support, patience and guidance that my wonderful advisory team has provided me — Dr. Emma Westecott, Dr. Adam Tindale, and Dr. Steve Engels. You were so elemental to this project, and to my discovery (and rediscovery) of new fields.

Thank you to my family for being so patient with me when I had to spend my nights shut away in my room working on my thesis, and to my partner Ritik for nurturing my passion for HCI further. As always, thank you to my dad for pushing me to get my masters and for sacrificing so much to make my life what it is.

To my sweet Lucky, though I lost you on the way, you have a paw print on my heart forever.

Contents

List of Figures

1 Introduction

Verbal toxicity in online multiplayer games is a pervasive issue that affects both players and game development companies. Approximately 75% of online players in the United States experienced harassment in 2023, amounting to about 110 million players in total, including young players from ages 10 to 17 years old [1]. The prevalence of toxic behavior in games has led it to become normalized across both games and generations of players since younger players learn to exhibit the same behavior [1]. Many players turn to video games for enjoyment, but instead toxicity manifests distressful and harmful environments that impact victims' mental health and self-identity. About 20% of players also spend less money on online games due to the toxicity they experience [1], which results in a significant loss of profits for developers and poses an additional incentive for change. Players become deterred from toxic gaming spaces, which reduces their engagement in game communities and development, thus reducing their representation in games and likelihood of the issues being resolved, forming a harmful feedback loop that needs to be broken (Figure 1.1).

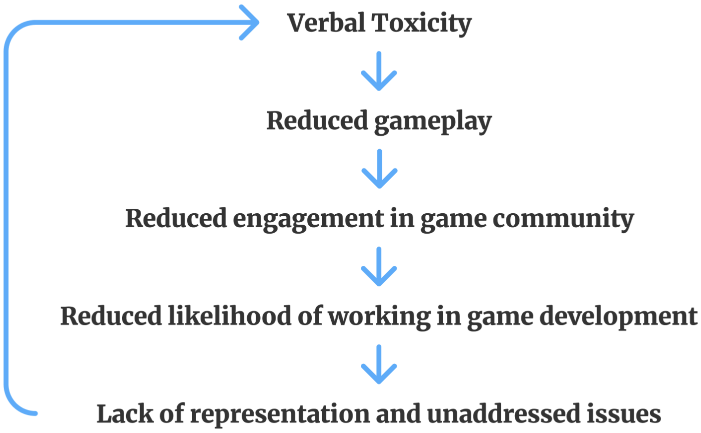

Verbal Toxicity
Reduced gameplay
Reduced engagement in game community
Reduced likelihood of working in game development
Lack of representation and unaddressed issues

Figure 1.1: Harmful feedback loop caused by in-game toxic speech.

Video game companies should play an active role in fostering respectful and safe environments for all players due to their unique ability to implement direct in-game solutions. While toxicity can stem from both player behavior and game design choices, developers have the greater capability to enact a form of governance for toxicity. Games have the ability to empower players, offer community engagement, and provide solace [1, 2]. The motivation of this research is to improve the safety and equitability of gaming environments such that a wider range of players can reap these benefits.

Toxicity is a complex term, but for this thesis, the definition provided in the *Disruption and Harms in Online Gaming Framework* by the Anti-Defamation League (ADL) and Fair Play Alliance [3] is used, as it offers standardized definitions based on a comprehensive review of existing terminology in this problem space. The framework defines **toxicity** as a blanket term encompassing any concerning or unacceptable behavior from a player or company, which can manifest as either disruptive behavior or harmful conduct [3]. **Disruptive behavior** is defined as any action that “does not align with the norms that a player and community have set” and consequently diminishes player experience and community well-being [3]. **Harmful conduct**, on the other hand, refers to “behavior that causes significant harm to players” [3]. Building on the definition of toxicity, **verbal toxicity** in this thesis will refer to spoken hate, harassment, or disruptive speech between online players in gaming voice chats. Toxicity is exhibited through various channels, including text chats, in-game actions, and more. Due to limitations in scope, this thesis focuses solely on verbal toxicity.

To address the issue of verbal toxicity in video games, the research question of this master’s thesis is: *how can design thinking approaches and natural language processing techniques be leveraged to build an intervention system that effectively detects, moderates, and potentially prevents toxic verbal discourse in online multiplayer games?* This thesis outlines the creation of such an intervention system that could be adapted and embedded into video games in the future. The intervention system estimates toxicity of speech through analysis of both text transcription and audio features, then triggers an appropriate on-screen response when necessary.

The backend processing for toxic speech detection and analysis was developed using natural language processing (NLP) techniques. **NLP** is a branch of artificial intelligence used for algorithmic processing, analysis and representation of human languages [4]. In this thesis, NLP was conducted through deep learning using artificial neural networks to conduct nuanced automated learning. Two core NLP algorithms were developed using speech audio as input; one for text-toxicity analysis and one for audio-toxicity analysis. The text-toxicity analysis algorithm transcribes the speech audio into text and analyzes the toxic content of the words. The audio-toxicity analysis algorithm instead looks at the audio features of the speech, such as pitch and volume, to analyze tone. The refinement of the machine learning algorithms was aided by the contributions of Yun Seok Yang, an undergraduate student in the Engineering Science department of University of Toronto, also under supervision of Dr. Steve Engels.

Audio moderation is less prevalent and researched than text moderation due to its increased nuance and complexity [5]. Historically, NLP has been more focused on text, largely because of the abundance of text-based data [4] and the challenges associated with audio analysis. Specifically, audio analysis requires significantly more computational power and large, multidimensional datasets [5]. Major companies such as Reddit and Twitch have implemented active measures in text moderation using software robots (‘bots’), third-party applications, and volunteer moderators to monitor discourse [6]. While these approaches demonstrate that toxic speech prevention techniques are widely used and supported, they have not been extensively researched or precisely applied to voice-based discourse. Factors such as tone and sentiment can alter the intent behind words, shifting the perceived toxicity of a phrase. Analyzing an extra modality of verbal toxicity through audio features has the potential to greatly enhance the accuracy of toxicity detection.

Deliberate design thinking was used to determine the types of interventions to initiate in various scenarios, and then to design interface elements that enact these interventions and that integrate seamlessly into gameplay. Design thinking was chosen as a methodology due to its emphases on empathy and iteration when solving complex problems. Together with the backend processes, the components and flow of these processes are shown in Figure 1.2.

Two major uncertainties arise from the flow diagram in Figure 1.2. Firstly, which audio

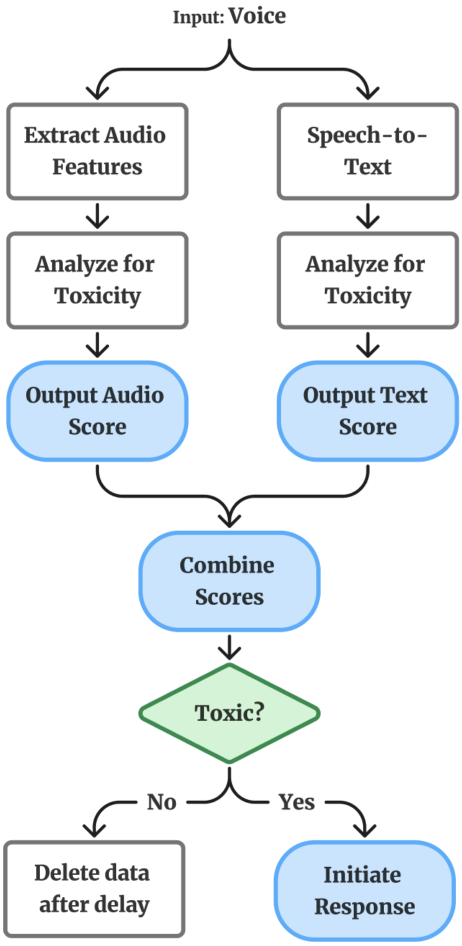


Figure 1.2: Flow diagram of the main components in the intervention system. Grey rectangles represent actions, blue rounded rectangles represent outputs, and the green diamond represents decisions. Yang was involved in the backend development of the system (everything before the green diamond).

features are measurable, and can effectively inform the toxicity levels of speech? Audio feature extraction for toxicity is under-researched and thus, this research attempts to make pioneering efforts to discover which features could prove most effective. Secondly, how should the on-screen intervention system be designed such that it responds to toxicity in a manner that optimizes for player-system cooperation, player experience, and correction of toxic behavior? These conditions for optimization were chosen based on literature review findings on the main failures of existing moderations systems, such as players showing low usage of reporting functions, finding moderation methods obtrusive, and lacking accountability to correct toxic behavior [1, 7].

To fulfill the optimization conditions, both the interface elements and interventions were designed by referencing a contextual review of existing systems, performing user experience heuristic evaluations and collecting player testing findings. Hochleitner et al.'s [8] framework of user experience heuristics for games were referenced for heuristic evaluation, particularly in the areas of learning, feedback, visual appearance and interaction. Ultimately, the three optimization conditions and the heuristic evaluation were the determining criteria for the success of the intervention system.

To optimize for correction of toxic behavior, operant conditioning techniques were applied. **Operant conditioning** refers to the process of behavioral learning through consequence outlined by psychologist B.F. Skinner in 1937 [9]. Specifically, operant conditioning involves the change of a reversible behavior through the delivery of some stimulus according to a well-defined rule [9]. The two methods of operant conditioning are given by reinforcement, which encourages a behavior, and punishment, which discourages a behavior [9]. These methods can be applied by either providing something ("positive") or by taking something away ("negative"). For instance, positive and negative reinforcement both encourage a behavior by providing a pleasant stimulus and taking away an unpleasant stimulus, respectively [9]. On the other hand, positive and negative punishment both discourage a behavior by providing an unpleasant stimulus and taking away a pleasant stimulus, respectively [9]. Typically, many intervention systems rely on negative reinforcement after a toxic incident has occurred, which does not effectively correct and prevent toxic behavior [10]. To optimize for the correction of toxic behavior in this intervention system, efforts were made to incorporate more positive reinforcement to encourage friendly behavior.

2 Literature & Contextual Review

The literature and contextual examples reviewed for this research were selected to gain a thorough and comprehensive understanding of the mechanics of verbal toxicity, or toxicity in general. Toxicity in games is a prominent topic in contemporary game research due to its widespread prevalence in gaming communities and the drive to improve moderation methods [5]. Due to these aspects, toxicity is being studied from an increasing number of disciplines, including computer science, social psychology, human-computer interaction, ethnography, and more. Accordingly, resources from these disciplines were reviewed to contextualize toxicity and to examine existing intervention systems for toxicity moderation.

2.1 Framing the Problem Space – Toxicity

To effectively address verbal toxicity in online games, it is essential to first examine and understand its complexity. Toxicity in online games, or more broadly in online spaces, can be envisioned as a dynamic cloud system that intersects areas such as identity, power dynamics, and skill level (Figure 2.1). These 'clouds', representing the boundaries of toxicity, can shift unpredictably, influenced by time and current events. Perceived toxicity varies from person to person given its subjective nature, resulting in fluid and non-discrete boundaries. At the same time, there are delicate borders between what is considered toxic and non-toxic behavior within a common context. For instance, discriminatory remarks targeting an individual's identity can be considered toxic, but peaceful conversation about identity should be acceptable. Visualizing toxicity as a cloud system can shed light on its complexities, and raises the critical question: *how can we effectively define and outline its boundaries?* To address this question, the following section synthesizes the reviewed literature using the 5Ws framework following a design thinking methodology, as suggested by Ambrose & Harris [11]. The 5Ws framework—which examines Who, What, When, Where, and Why—offers a structured approach to clarify this complex topic.

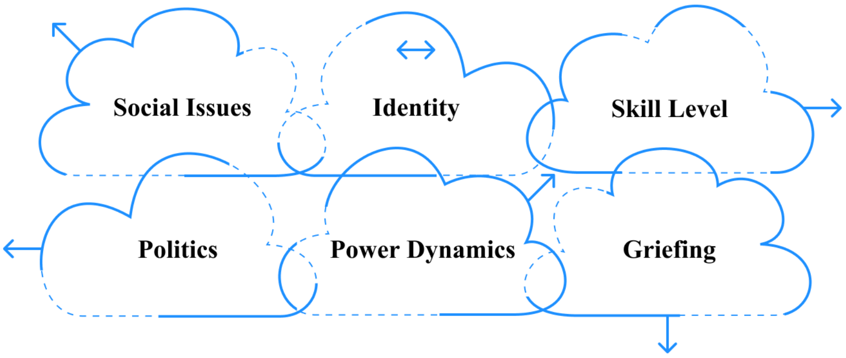


Figure 2.1: Dynamic cloud system analogy for the complex boundaries of toxicity. These are some of many topics of verbal harassment in online games [1].

2.1.1 Who – Target Players

When a case of toxic speech occurs in an online multiplayer game environment, players generally find themselves in one of three main positions during an instance of toxicity: the **victim**, the **aggressor**, and the **witness**. The victims are the targets of the toxic speech, the aggressors engage in toxic speech, and the witnesses are not directly involved but can observe the toxicity by being present in the same space. The subsections below aim to propose common features of each of the aforementioned positions.

Demographics of Victims

Given that online gaming communities are very diverse, there are players of various demographics that have differing experiences in gaming spaces. To gauge the unique experiences of player demographics, the international non-governmental organization Anti-Defamation League publishes annual reports on hate and harassment in online games and details the experiences of specific communities. For the report published in 2023, the game data platform Newzoo had conducted player surveys with 1,971 participants aged 10 to 45 years old, as well as 10 qualitative interviews [1]. The main topics that were covered in the surveys and interviews involved overall harassment experienced (by age and identity), sense of safety, specific games with high frequency of toxicity, and experiences of reporting other players. From this report, it was found that teens and preteens (aged 10 to 17 years old) experienced an increased amount of in-game harassment from 67% in 2022 to 75% in 2023 [1]. Among this group of young players, 37% experienced identity-specific harassment (up from 29% in the previous year) [1]. In terms of identity, women and Black or African American adult gamers experienced the most identity-specific harassment at 48% and 50%, respectively, up from 47% and 44% in 2022 [1]. Simultaneously, Jewish adults experienced the most harassment of any form at 70% of all players [1]. Being aware of which player demographics experience the most toxicity can bring insight to the design of the intervention system and the player testing process since it ensures that the maximal number of wants and needs are being considered. For instance, during the player testing of the prototype, a unique request was made by an individual to include certain racial slurs they have been victim to in the training dataset for the backend toxicity detection algorithms. They had realized that the system did not flag the use of certain slurs since they are missing from the dataset. The datasets used can introduce their own biases, so welcoming input from a diverse set of players can help to account for these overlooked areas and allow for a more equitable player experience.

Women in gaming communities have historically faced high levels of sexism and misogyny, a reality that many players recognize as a persistent issue. ADL [1] notes that in all of their previous annual surveys, women remained as the group that experienced the most harassment. Sexism toward women was also a recurring topic during player testing of this thesis prototype, with participants of all identities acknowledging the issue. Female gamers are stigmatized by society, which induces psychosocial harm and high levels of stress during gameplay against men [2]. To avoid harassment, women tend to resort to playing offline, anonymously, non-verbally, or with voice-masking [2]. Unfortunately, these problems lead to hostility even between female players, which further increases exclusivity and invalidation.

Ethnographic Work on Game Communities

Numerical data regarding the toxicity experienced by various demographics is tremendously useful and has been widely studied since the emergence of virtual spaces. More recently, ethnographic research has been introduced into the field of game research to provide insights into the specific sociological phenomena that players from different cultures or communities may experience in online gaming. It can be valuable to include this more critical, equity-facing approach that studies specific gamer communities in more depth, which is often missing from human-computer interaction and computer science approaches. Dr. Kishonna L. Gray's work acts as an important reference as the first widely-cited Black feminist researcher working at the intersection of gaming and ethnography. The following sections explore insights from Gray's book *Race, Gender, and Deviance in Xbox Live: Theoretical Perspectives from the Virtual Margins* pertaining to deviant behavior within the gaming community of Xbox Live [12].

Deviant behavior – Gray [12] uses the term **deviance** to refer to behaviors that are non-conforming to the unspoken etiquette and rules of online game spaces. This definition matches the aforementioned definition of disruptive behavior from the *Disruption and Harms in Online Gaming Framework* by ADL. Some examples of deviant or disruptive behavior include rude comments made by certain players that expect a more serious play style than others [3], or yelling slurs to trigger others. The stereotypical players–being male and experienced–remain the dominant community capable of perpetuating **deviant behaviors** and discriminatory norms typically towards marginalized and novice players. Gray [12] points to the profound contradiction in how this seemingly dominant group can exclude and alienate minority groups which actually constitute a significant portion of the gaming community numerically. Deviance can result in othering of marginalized players by creating an atmosphere that attempts to exclude certain groups and reducing their sense of belonging, which stems from the inherent human need for belonging to a group with a similar sense of identity [12]. In voice chats specifically, aggressors may use audio cues from their victims to comment on the victim's identity, essentially profiling them and invading their privacy [12]. Despite not constituting a numeric majority, these aggressors maintain their status by engaging in toxic behavior to exert dominance over other groups, which can demotivate victims from improving their gameplay performance and diminish their enjoyment of the gameplay. Acts of dominance can foster a culture that normalizes toxic gameplay, and that reinforces deviant behavior.

Online Disinhibition as a Cause – Deviance online has been theorized to be a consequence of the online disinhibition effect, coined by Suler and Phillips, as cited by Gray [12]. The **online disinhibition effect** proposes that anonymity is the main factor that facilitates the emergence of racist and sexist behaviors in online spaces [12]. From a social psychology perspective, anonymity has been found to provide users of the internet many benefits to their psychological wellbeing through factors such as privacy and autonomy [13]. Privacy allows for users to have a presence on the internet without disclosing personal information that could otherwise be taken advantage of [1]. Autonomy gives users the opportunity to "... experiment with new behaviors without fear of social consequences," or being identified by those they know [13]. These factors can provide users with a sense of freedom and safety when engaging in computer-mediated communications. However, anonymity and lack of restraint also allows for minimal social accountability or consequences, enabling individuals to express prejudice and hostility they might otherwise suppress in physical spaces.

The online disinhibition effect can also be attributed to several psychological phenomena beyond anonymity and lack of restraint. Many players, and users of online spaces in general, tend to treat virtual environments with different mindsets and behavioral standards than they would use in physical spaces. This treatment of the virtual world can be described by the phenomena of solipsistic introjection, dissociative imagination, and lack of authority [12]. **Solipsistic introjection** describes the situation where aggressors conceive or perceive the voice or face of another user due to the user's lack of descriptive visual or communicative cues that would otherwise develop their identity in cyberspace [12]. Solipsistic introjection can be harmful since the common practice of keeping identity private online can result in an aggressor perceiving an identity that the aggressor deems is inferior to themself. Victims have to resort to defensive measures such as muting their microphone, silencing audio, or exiting public voice lobbies. These measures can pose a disadvantage since they must sacrifice their ability to hear sound effects and communications for team-based strategizing that would otherwise benefit their performance and enjoyment in the game. Some female players also rely on voice altering technologies to use a deeper voice. On the other hand, **dissociative imagination** occurs when players perceive virtual spaces to be imaginative disembodied spaces that are separate from physical spaces and regular social etiquette. Dissociative imagination is more common in fantasy games and can result in aggressors devaluing the virtual space, as well as expecting victims to devalue their toxic behavior. Finally, the **lack of authority** in cyberspace refers to the lack of a guardian or form of supervision to prevent deviant behavior and to establish the status of an aggressor after they commit toxic behavior [12]. Even if an authoritative figure is placed in these spaces, they would not be able to sustain much power due to the lack of physical presence and cues [12]. These phenomena can provide clarity on the cause of mistreatment of virtual spaces and the need for some form of intervention system that supervises and interrupts deviant behavior.

Game Narrative Influence – Unfortunately, the cause of toxicity in games is multifaceted, and is exacerbated by the prejudices that arise not only from players, but also the ones that are ingrained into the games by developers. Video games can represent the developers' values, which can in turn influence and impact players. This influence can enable or incite toxicity as well. Gray [12] touches on how game narratives can perpetuate and reinforce ideologies about social structures and power dynamics, all while disguising them with entertaining storylines that effectively normalize the ideologies. She also provides specific contextual and historical examples to demonstrate the issues of racial and gender-based discrimination in the video game community.

The widely popular video game series, *Grand Theft Auto (GTA) V* [1] presents one example of many. GTA engages players in mass crime and is a satire of American popular culture in a way that reinforces the privileges of masculinity and whiteness [12]. GTA explicitly displays scenes of violence and hypersexualization of women, and of racial stereotypes within the main characters. Exposing players to these racialized and sexualized narratives affirms the status quo regarding racial stereotyping and can result in players carrying over these behaviors into their interactions with real humans [12].

Player Archetypes

Players can be generally understood through common archetypes that have been established by various researchers over time. Prior to conducting player testing, identifying the archetypes participants relate to can help gauge how well they represent different player types. This insight can be used to ensure that the wants and needs of a diverse set of players are taken into account when designing the intervention system. Within a human-computer interaction research study in gaming by Caci and Dhou [14], it was reported that personality is integral to determining "... why and how people are immersed in playing games," and that players typically play games that align with their personalities, worldviews and identities. For instance, players who are extroverted may prefer games with more social features, and players who are more conscientious may value games that involve strategic thinking [14]. However, it is important to keep in mind that while considering archetypes, it cannot be assumed that all players can fit within just one or even any of the named categories. Players may also switch from one archetype to another based on the specific game, game genre, mood and playstyle [12].

An early model outlined by game researcher Richard Bartle [15] provides four player types that are primarily based on player motivations, in-game behaviors, and play styles. However, it must be noted that these archetypes were established for players in multi-user dungeon games (MUDs), that primarily entail text-based role-playing. As a consequence, these archetypes provide less accuracy for players of other genres, but they are still studied as they are precursors for player archetype models that were established later. The following definitions are referenced from Bartle [15].

- **Achievers**: Inclined towards collecting and potentially displaying in-game achievements (i.e. points, status).
- **Socializers**: Motivated by positive emotions brought about from interacting with other players.
- **Explorers**: Driven by the curiosity to explore new environments, narratives, or other game elements.
- **Killers**: Motivated by winning and collecting achievements, but also find enjoyment in seeing other players lose and/or causing distress.

From this list, Gray [12] proposes that the Killer player archetype tends to engage in and influence the most deviance. One deviant behavior that is extensively studied in game research is griefing in first-person shooting (FPS) games. **Griefing** refers to a behavior where aggressors intentionally disrupt others by exploiting game mechanics in unintended ways for personal enjoyment [12]. An example of this is when aggressors kill their own teammates or destroy team structures so that they can feel a sense of power over others. The motivations of griefing can be connected to the motivations of the Killer archetype since they both have an overlap of disrupting other players for pleasure. This overlap suggests that it is worth further investigating connections between player archetypes and toxic players.

Bartle's four player archetypes were expanded into seven archetypes by Bateman and Boon in 2005. These expanded archetypes are listed below, as cited by Caci and Dhou [14].

- **Achievers**: Same as Bartle's Achiever archetype.
- **Socializer**: Same as Bartle's Socializer archetype.
- **Seeker**: Same as Bartle's Explorer archetype.
- **Survivor**: Drawn towards fictional scenarios that elicit a sense of fear and danger.
- **Daredevil**: Thrives off thrill, risk-taking and adrenaline.
- **Mastermind**: Enjoys puzzle games and thinking strategically.
- **Conqueror**: Moved by overcoming challenges and beating opponents.

Due to the increased nuance and number of archetypes provided, Bateman and Boon's model can offer better representation for a larger portion of a gaming community or player base. A notable feature about Bateman and Boon's model is the lack of the Killer archetype from Bartle's model. It can be speculated that instead the Achiever, Daredevil and Conqueror categories might give rise to deviant behaviors since they are focused on overcoming challenges or players, and do not encourage much collaboration. Alternately, Socializers would require healthy relations to other players. Seekers, Survivors and Masterminds would be more focused on game mechanics that might require or also be improved by collaboration.

To add nuance to the discussion of different player types, examining broader personality traits and behaviors beyond game contexts can offer valuable insights. The **Five Factors Model** is a widely known model that provides five dimensions of personality: Extraversion, Openness, Conscientiousness, Neuroticism, and Agreeableness. The five factors are considered to be dimensions since an individual can be placed on a spectrum of how much they identify with the traits of that factor. The traits of these five factors are provided below, as cited by Caci and Dhou [14].

- **Extraversion**: How energetic, assertive, and sociable they are.
- **Openness**: How creative, curious, and imaginative they are.
- **Conscientiousness**: How disciplined, organized, and ambitious they are.
- **Neuroticism**: How easily they can become angry, anxious, and depressed.
- **Agreeableness**: How compassionate, trusting, and cooperative they are.

Having these high-level categories of personality can add an extra layer of depth when conceptualizing the relationships between player type and tendency toward deviant behavior.

2.1.2 What – Forms and Modalities of Toxicity

Aside from understanding the player interactions in gaming spaces, it is also important to identify the various ways toxicity materializes. Toxicity can be considered to primarily exist in the form of hate and harassment [1]. In ADL's 2023 report, it was found that harassment largely occurs through identity-based harm, but also through extremism and misinformation towards 15% of adults around themes of white supremacy, anti-LGBTQ+ rhetoric and antisemitism [1]. Conversations revolving these topics are typically incited by current events such as presidential elections and international tensions. Other reported forms of harassment, ranked by frequency, include name-calling, trolling, griefing, sustained harassment, threatening, sexual harassment, doxing, and swatting [1]. Being aware of the forms of toxicity is critical in determining the mechanics of the issue, so that they can be appropriately mitigated.

Within video games, toxicity is widely known to primarily occur through voice and text-based communication. However, toxic behavior can also be demonstrated through in-game actions such as griefing and trolling [12]. Another channel of toxicity is through external social platforms where gaming-related discourse takes place. An example of toxicity through social platforms is the Gamergate incident, which was a misogynistic online harassment campaign that took place in 2014 [16]. This incident began when an ex-partner of a female game developer, Zoë Quinn, circulated an article making false allegations of Zoë getting her job position through sexual favors [16]. The article initiated online misogynistic outrage from players who felt threatened about the increasing number of female players and developers in the game community [16]. This led to online abuse and harassment towards Zoë, any women who showed support for Zoë, and the families of these women. The campaign involved hacking the women's accounts and leaking information (defined as **doxxing**), death and rape threats, stalking, and more [16]. Gamergate demonstrates that there are many avenues for toxicity to take place. Though these are important channels of toxicity to consider, they are not within the scope of this research endeavor. Based on ADL's 2023 report, voice-channels have the highest frequency of toxicity, yet are under-researched, making this research both timely and essential.

2.1.3 When and Where – Genres and Games

The frequency and type of toxicity players experience largely depend on the gaming environment, with unique patterns emerging across specific genres and games. ADL has collected data on sense of safety and frequency of reported harassment in certain online multiplayer games and has theorized that competitiveness and game pace are key drivers in toxicity levels [1, 12]. Less competitive and slow-paced games like *Minecraft* tend to be safer, with 40% of adults reporting toxicity in 2023 [1]. In contrast, more competitive and fast-paced games like *Call of Duty* and *Dota 2* saw 83% and 88% of adult players reporting toxicity, respectively [1]. An intriguing finding from ADL is that players often recounted a similar sense of safety to express themselves across these games, suggesting that degrees of competitiveness may be a critical factor that flips the intent for harm behind player expression. While competition is integral to many games, understanding its role in toxicity can provide insight on how mitigation strategies should be tailored to games.

A particular trend in the modern game industry is the growth of live service games, which are games that have periodic releases of new content on a subscription basis, such as a season pass. A primary incentive for game development companies to release life service games is the continuous revenue that it offers [1]. A consequence of this structure is that it allows for the player base to develop their skills over time, resulting in highly skilled players and the development of a certain culture for that game. Unfortunately, this can often lead to the developers being more accommodating to the wants of long-time players, resulting in a difficult onboarding experience for novice players, and in higher levels of discrimination from long-time players toward novice players. This is evident in live-service games such as *Destiny 2*, which has had active releases since 2017. By consistently providing new content to fans, the game world and experience has gotten richer, but it has also resulted in a steep learning curve for novice players. This knowledge and skill disparity between novice and long-time players can result in increased verbal harassment. Consequently, it can be found that aspects such as the delivery method of gaming content can also impact the toxicity levels in a game community.

2.1.4 Why – Incentives for Change

For players – In 2023, there were about 3.38 billion video game players worldwide and the game development industry had amassed $184 billion USD [1]. Exposure to toxicity can diminish enjoyment, sense of community, and performance, while increasing stress [7]. 76% of adult online players being affected by toxicity amounts to about 2.57 billion players experiencing or witnessing harassment. It is clear that toxicity has a negative impact on the well-being and entertainment of a significant portion of players, and the negative consequences on the developers of video games can be made clear through direct observation of player spending habits.

For development companies – A profound finding from ADL's 2023 report is that 20% of players make less purchases in online games after experiencing toxicity [1]. In-game purchases account for 27% of console games' revenue and 97% of mobile games' revenue, totally to $15.3 billion USD and $89.7 billion USD, respectively. These figures highlight the vast potential for revenue loss directly from toxicity [1]. ADL's [1] qualitative interviews with 10 players revealed that most respondents play for entertainment, but the harassment they experience diminishes their enjoyment. Some respondents shared that they associate a game with the harassment they experience, and that they feel as if both the game and its community represent the same ideologies as the aggressors [1]. One respondent stated that toxicity "'changes whether we see [spending money on games] as worthwhile or not. [It] definitely will make an impact on what we want to spend. Sometimes we just don't want to throw our money into something that we don't enjoy as much'" [1]. Negative associations of a game can diminish retention of existing players, but also acquisition of new players, creating a feedback loop [7]. Investing in inclusive, well-designed content moderation is critical to fostering engagement and maintaining revenue.

2.2 The Resolution – Intervention Systems

Following the investigation of the 5Ws of toxicity, it is natural to question the 'How?' [11]: how will this problem be solved? To resolve the complex issue of toxicity for a diverse set of players and game developers, it is necessary to be informed on these concepts and statistics on toxicity, but also to reflect on the research and interventions that have been implemented so far.

2.2.1 Designing the Intervention System

From a systematic literature review that was conducted on 36 intervention systems for video game toxicity by Wijkstra et al. [7] in 2023, there were three main conclusions that were found; 28 of the systems introduce new approaches rather than iterating on existing ones, 31 of the systems only intervene after an instance of toxicity occurs, and only 5 of the systems used data from commercial games or platforms. These conclusions suggest that there is a lack of iteration upon existing intervention systems, that more approaches should be taken to prevent toxicity altogether, and that the systems should use commercial data and be tested with players to promote external validity [7]. External validity here would be measured through the optimization conditions of player-system cooperation and experience. To employ these insights in the design of the intervention system for this thesis, existing examples are analyzed and prioritizing correction of toxic behavior. Potential toxicity prevention techniques include the establishment of legal agreements or policies for game development and gameplay. Unfortunately, there is minimal access to commercial game data for this project, but player testing is used.

2.2.2 Existing Interventions – Contextual Review

Researchers Yang et al. [17] at the game development company, Ubisoft, have developed a deep learning model using BERT architecture and 194000 text-chat phrases from their games, *Tom Clancy's Rainbow Six Siege* and *For Honor* [17]. Both games offer online multiplayer game modes, but Rainbow Six Siege requires individual skill, while *For Honor* involves team strategizing [18]. Involving different forms of multiplayer games can shed light on how differing player dynamics can affect the amount of toxicity present. Yang et al. achieve model performance metrics of 82.95% in precision, and 83.56% in recall, which are 7% and 57% higher than pre-existing models, respectively [17]. Using commercial game data, Yang et al. demonstrate the importance of collecting game-specific language when training a model for game toxicity moderation. They report the improvement in the model's ability to identify reported players, but also the transferability of this performance to other similar games [17]. Context is another crucial indicator in determining the toxicity of a phrase, as Yang et al. emphasize, despite its frequent omission in moderation research due to the challenges of detecting context. They introduce the first chat toxicity detection model capable of analyzing context and demonstrate improved performance as a result. They experimented with Jigsaw's Toxic Comment Classification Challenge data by providing context for only half the dataset and observed an increase in toxicity annotation from 4.4% to 6.4%, suggesting that context can offer high accuracy in the detection of toxicity. Overall, this computational research demonstrates the necessity to use game-specific data, especially from commercial games, to improve external validity of the intervention system. However, the lack of audio or visual data leaves many cases of toxicity unaddressed. Yang et al. propose that future interventions include real-time toxicity alerts, educational programming targeted towards players, and behavioral studies to understand the outlook of player interaction with intervention systems [18].

In Gray's [12] review of Xbox Live, she analyzes Xbox 360's introduction of gamer zones to target verbal harassment. There were four gamer zones: Recreation, Family, Pro, and Underground. These zones were established to allow players to label themselves in certain skill levels and intentions, and Family was used to encourage family-friendly language [12]. Unfortunately, the gamer zones were not enforced in matchmaking and did not affect gameplay, so this system was not successful and did not provide any meaningful solutions to the core issue [12]. This example highlights the need for a more rigorous and controlled system with clear rules that cannot be overstepped and involves reinforcement.

A related commercial intervention system is ToxMod, created by the start-up Modulate. ToxMod was introduced in 2020 and has been developed into a web plugin that uses advanced machine learning to moderate voice chats for games in 18 languages [19]. ToxMod can assess the tone, timbre, emotion, and context of a conversation to determine the type and severity of toxicity [19]. Modulate's main priorities are player safety and privacy, and this tool outputs information in the form of flagged cases and statistics to moderators (Figure **??**). This system sets a strong example of toxicity detection and was found to be successful for Activision in 2024. ToxMod provides clear, integrable solutions that can further incentivize game development companies to employ such a tool and acts as a strong reference.

2.2.3 Reporting

Literature on existing intervention systems indicates that user experience has not been extensively considered, especially in reporting. Only about one third of victims ages 10 and up report toxicity since they find it tedious, feel that it is now ingrained into the online gaming experience, and do not want to be deemed as "snitches" [1]. ADL's [1] 2023 report specifically suggests that two of the best solutions to resolve toxicity in-game would be to have more accessible and efficient reporting systems, and to strengthen moderation of voice chats which trails that of text chats [1].

3 Methodology

Given that verbal toxicity in online multiplayer games is a complex and sensitive issue, a delicate approach must be used to address it.

3.1 Design Thinking

To frame the research process, **design thinking** was used since it allows for innovation through a human-centered and iterative approach. Different forms of design thinking exist, and the best option depends on the nature of the problem and the desired approach. The Hasso Plattner Institute of Design at Stanford University [20] outlines 5 phases: Empathize, Define, Ideate, Prototype, and Test. These phases are referenced in Figure 3.1, along with their potential applications in this project, and are not necessarily followed linearly. The Ideating, Prototyping and Testing phases were repeated throughout the research process. It is important to note that this interdisciplinary research was conducted at the intersection of human-computer interaction, machine learning, game studies, ethnography, social psychology and equity research. Approaching the project from these fields impacts how the design thinking phases are perceived and executed. In this case, Empathize and Define were considered in respect to players, developers, and their experiences with toxicity. On the other hand, the Ideate, Prototype and Test phases were applied for the iterations of the intervention system.

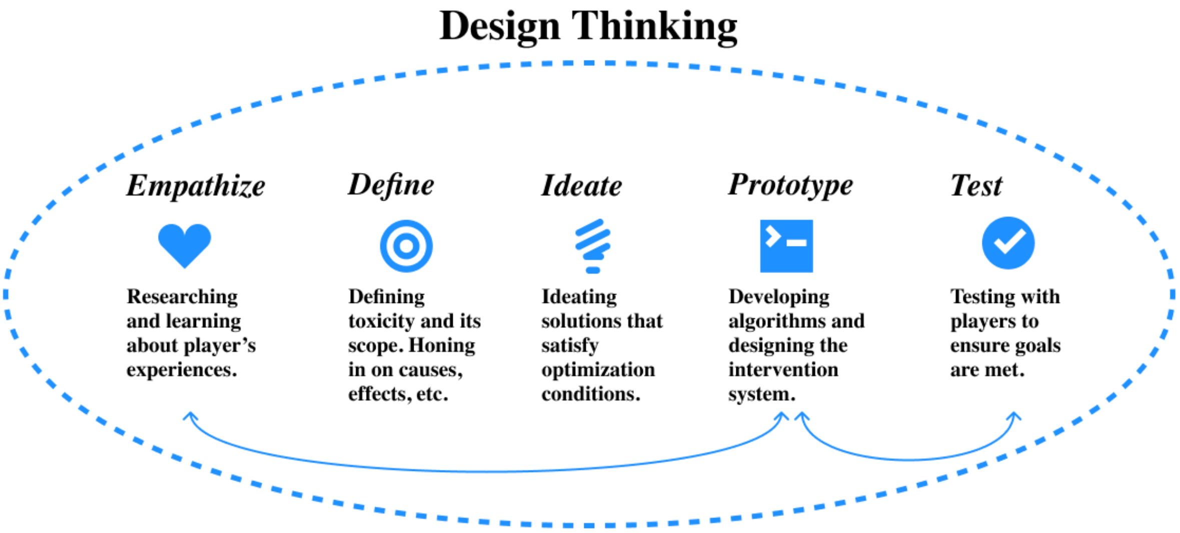

Design Thinking
Empathize
Researching and learning about player's experiences.
Define
Defining toxicity and its scope. Honing in on causes, effects, etc.
Ideate
Ideating solutions that satisfy optimization conditions.
Prototype
Developing algorithms and designing the intervention system.
Test
Testing with players to ensure goals are met.

Figure 3.1: Design thinking phases, as outlined by the Hasso Plattner Institute of Design [20] and adapted to resolving in-game verbal toxicity for this thesis.

Empathize involves conducting user research to gain insights about the user's experiences [20]. In the case of this research project, empathizing with the victims of toxicity is a crucial step to begin addressing the problem, as well as the players' wants and needs. Empathizing was primarily executed through review of existing literature (Section 2) and player testing results

(Section 4.2.4). For instance, learning about the unique problems that certain demographics experience in gaming spaces through ADL's annual player surveys [1], or hearing about individual player needs during one-on-one conversations through player testing can build the empathy required to develop a sensitive and accommodating intervention system.

Defining shifts the focus towards understanding the mechanics of the problem [20]. Toxicity is undoubtedly a complex issue due to its subjective and shifting nature. When designing an intervention system that detects toxicity, the boundaries of toxicity as indicated by the dynamic cloud system analogy in Figure 2.1 must be established, or at least approximately determined. Understanding these boundaries are necessary to determine what is and is not considered toxic. The defining step was also done through literature and contextual review in Section 2.

Ideating involves brainstorming innovative solutions to the problem [20]. The priorities for the intervention system are to optimize for player-system cooperation, player experience and correction of toxic behavior. Ideating involved designing solutions that fulfill these conditions, as well as the chosen user experience heuristics. The ideating step was informed by the contextual review of existing intervention systems in Section 2.2, and was executed through use of storyboarding, user experience heuristics and user flows, as shown in Section 4.2.

Prototyping refers to the iterative development of the solution [20]. In this case, the iterative development was conducted for the machine learning algorithms that detect and measure the toxicity of speech, as well as the interface elements and interventions of the intervention system. The prototyping phase is shown in Section 4.

Testing evaluates the success of the prototype in resolving the problem and fulfilling the outlined goals [20]. In particular, the intervention system needs to be tested to determine if it accomplishes the goals of detecting, moderating and preventing toxicity, but also the optimization conditions. The testing step was done through player testing, and collection of feedback during the exhibitions of the prototypes. Details about the outcomes of testing are provided in Section 4.2.4.

3.2 Methods

The main methods that were used include interdisciplinary desk research, storyboarding, user flows, iterative development and design, user experience heuristic evaluations, user research, and thematic analysis through affinity diagramming. As shown in the literature and contextual review, references were pulled from a diverse set of sources to address the complex nature of toxicity, and to maximize external validity towards different demographics, games and scenarios. Findings were used to inform the creation of storyboards and user flows to *Ideate* the design of the system. Iterative development and design were conducted in accordance to design thinking phase *Prototyping* to progressively improve and refine the system. The development of backend processes involved several iterations through testing various combinations of datasets and feature extraction. The design of the interface elements and interventions went through three iterations, from a wireframed prototype, a live three-dimensional game prototype for testing, and refinement post-testing. User research was conducted with 10 participants through Research Ethics Board approval. Player testing was used to mitigate potential biases from both the designer and the underlying data, inform design decisions, gather feedback on methods of governance, and ensure that the system accommodates a diverse player base. The results of the player research were categorized thematically through affinity diagramming to identify common themes and overarching ideas.

User experience heuristic evaluations were used when designing the interface elements and interventions for the intervention system. Hochleitner et al.'s (2015) framework of heuristics for games were referenced from the areas of motivation, feedback, visual appearance and interaction [8]. These areas and their respective heuristics were chosen in accordance with the optimization conditions for the intervention system of player-system cooperation, player experience, and correction of toxic behavior. The specific heuristics that were chosen for use in this project are reproduced below [8]:

Learning

- 4.1 The player is given space to make mistakes, but the failure conditions must be understandable.

Feedback

- 8.1 The acoustic and visual effects arouse interest and provide meaningful feedback at the right time.
- 8.3 The feedback is given immediately to the player's action.

Visual Appearance

- 9.1 In-game objects are standing out (contrast, texture, colour, brightness), even for players with bad eyesight or colour blindness and cannot easily be misinterpreted.
- 9.2 Furthermore the objects look like what they are for (affordance).

Interaction

- 10.1 Input methods are easy to manage and have an appropriate level of sensitivity and responsiveness.
- 10.2 Alternative methods of interaction are available and intuitive. When existing interaction methods are employed, they are adhering to standards.

(Reproduced from [8].)

For data analysis, a mixed-methods approach combining both quantitative and qualitative research were applied. Qualitative research, conducted through player testing and literature review, provided insights into player experiences during gameplay. Meanwhile, quantitative research in the form of experimentation and optimization, guided the iteration of machine learning algorithms to enhance the accuracy of toxicity analysis.

4 Design and Development

The following sections outline the steps that were taken to build the intervention system, which involved developing the backend processes, and designing the interface and interventions. Both the development and design occurred concurrently, as shown in the overview of the timeline in Figure 4.1 below. Interconnections between these two areas are shown by the middle arrows, which indicate when the algorithms were implemented into the intervention system. Dashed borders indicate where Yang was involved. These steps were done in accordance with the design thinking methodology.

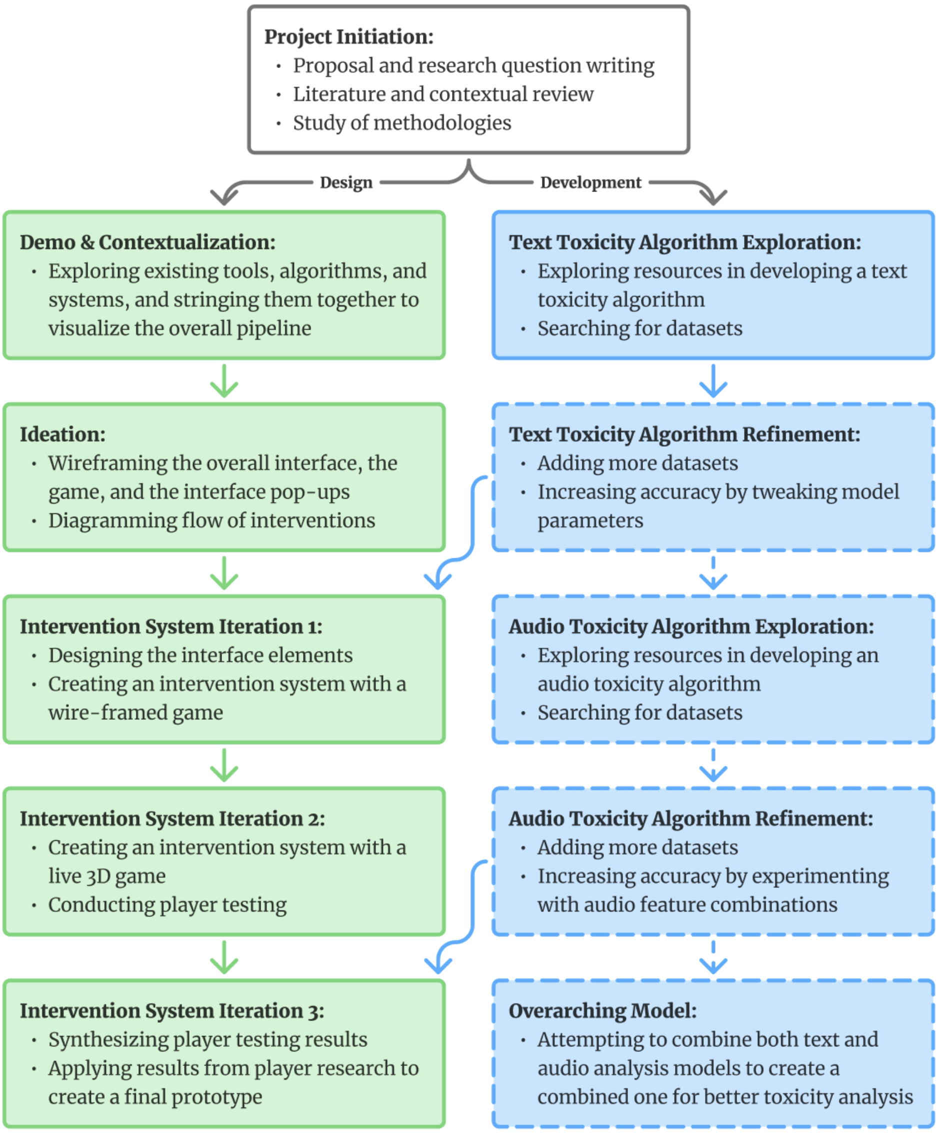

Project Initiation:
• Proposal and research question writing
• Literature and contextual review
• Study of methodologies
Design
Development
Demo & Contextualization:
• Exploring existing tools, algorithms, and systems, and stringing them together to visualize the overall pipeline
Text Toxicity Algorithm Exploration:
• Exploring resources in developing a text toxicity algorithm
• Searching for datasets
Ideation:
• Wireframing the overall interface, the game, and the interface pop-ups
• Diagramming flow of interventions
Text Toxicity Algorithm Refinement:
• Adding more datasets
• Increasing accuracy by tweaking model parameters
Intervention System Iteration 1:
• Designing the interface elements
• Creating an intervention system with a wire-framed game
Audio Toxicity Algorithm Exploration:
• Exploring resources in developing an audio toxicity algorithm
• Searching for datasets
Intervention System Iteration 2:
• Creating an intervention system with a live 3D game
• Conducting player testing
Audio Toxicity Algorithm Refinement:
• Adding more datasets
• Increasing accuracy by experimenting with audio feature combinations
Intervention System Iteration 3:
• Synthesizing player testing results
• Applying results from player research to create a final prototype
Overarching Model:
• Attempting to combine both text and audio analysis models to create a combined one for better toxicity analysis

Figure 4.1: Overview of the design and development timelines, and their interconnections. Yang was involved in all steps with dashed borders, and worked on the Overarching Model step individually.

4.1 Development of Backend Processes

To determine the toxicity levels of speech, the two modalities of text and audio were used. As shown in Figure 4.2 below, this involves a speech-to-text algorithm that transcribes the speech, a text-toxicity analysis algorithm that analyzes the toxic content of the words, and an audio-toxicity analysis algorithm that analyzes audio features to gauge sentiment and tone. These toxicity analysis algorithms are NLP models programmed in Python using the web-based interactive computing platform, Jupyter. The models were written using the TensorFlow deep learning framework, and are outlined in detail below.

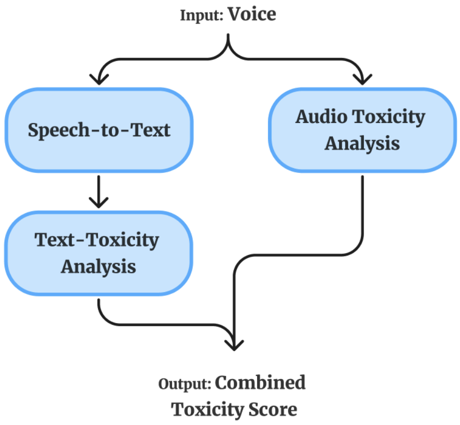


Figure 4.2: High-level overview of the development pipeline. The main three algorithms used in the backend are indicated by the blue rounded rectangles. Yang was involved in the refinement and training of all algorithms.

4.1.1 Text-Toxicity Analysis

The goal for the development of the text-toxicity analysis algorithm was to have a program that takes datasets of text phrases labelled with their respective levels of toxicity, trains a machine learning model using those pairings, and uses that model to measure the text-toxicity of any inputted phrase from outside of the datasets that were used.

To accomplish this goal, the Comment Toxicity tutorial in the form of a Jupyter notebook by Nick Renotte was referenced [21]. The dataset that was used for experimentation was Jigsaw's Toxic Comment Classification challenge dataset [22] composed of 153164 categorized Wikipedia comments categorized using six labels: toxic, severe toxic, obscene, threat, insult, and identity hate. This dataset was used because it offers a large amount of text-based data that aligns well with the goal of measuring toxicity.

Tokenization, a common preprocessing technique in NLP, was used to transform text phrases into a format suitable for machine learning models. Specifically, it split phrases into individual words, removed punctuation, and converted words into unique numeric identifiers called tokens. These tokens collectively form an encoded vocabulary. Each phrase was tokenized into a sequence of tokens, which were used for training alongside their corresponding toxicity levels. The tokenization process is illustrated in Figure 4.3.

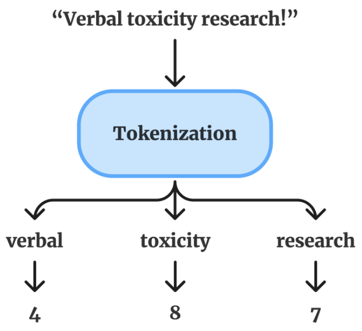


Figure 4.3: The process of tokenizing a phrase by stripping punctuation, converting to lowercase, splitting words, and assigning numeric codes.

Once the data was pre-processed using tokenization, it was passed into TensorFlow's Sequential API to form a linear stack of layers that form the model. The layers that were used included TensorFlow's Embedding, Bidirectional, and Dense layers as per the recommendation of the tutorial [21].

To assess the success of the model, performance metrics and player testing results were assessed. Multiple rounds of training were conducted to test the number of training epochs to use and which datasets to add such that the performance metrics, accuracy and recall of the model can be optimized (Figure 4.4). The final model was developed using datasets from Dota 2, Twitch and Youtube that provided a combined 36497 chat logs categorized into three levels of toxicity: non-toxic, somewhat toxic, and very toxic [23, 24]. These datasets were chosen since the phrases are from gaming-related, or similar platforms, and were categorized specifically for toxicity moderation. The final model provided an accuracy of 81%, precision of 71%, and recall of 71%. Compared to the precision and recall found by Yang et al. [17] of 82.95% and 83.56%, respectively, these results are lower, which is expected since Yang et al. [17] used a dataset of 194983 comments taken across three of Ubisoft's games and conducted more finetuning of the model parameters. There are limitations to this model, such as the fact that game-specific toxicity data is sparse.

Speech-to-Text

Since the text-toxicity analysis model requires text input, and players are providing speech input, there needs to be a step that converts the speech into text for it to be passed into the text toxicity model. Writing a speech-to-text algorithm from scratch can be an arduous task that is not within the scope or goals of this project. To find an existing tool for speech transcription, a comparative quantitative approach was used to test various pre-existing real-time speech-to-text machine learning tools for integration into this project. After testing various tools, success was found with the HuggingFace Speech Recognition API in Unity by Dylan Ebert. This tool was chosen over others because of its low error rate and compatibility with Unity, which is the chosen platform for the making of the intervention system.

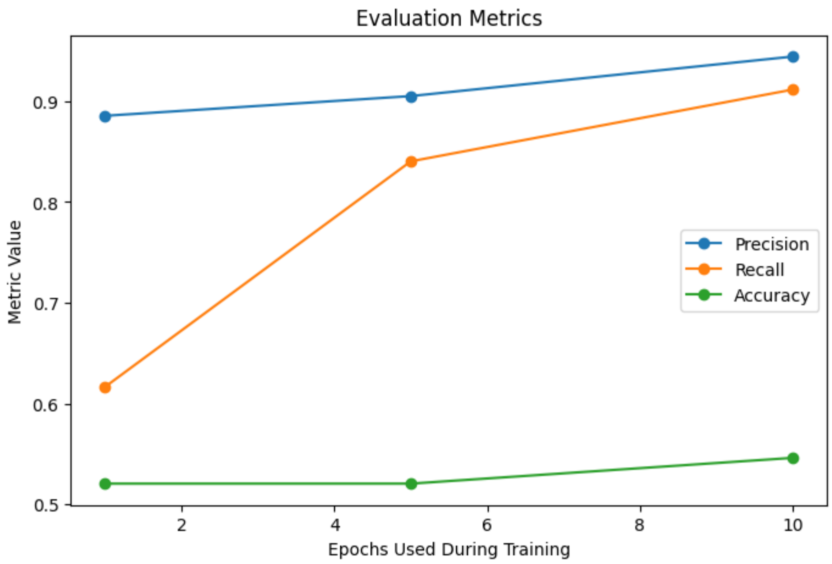


Figure 4.4: An example of one test that was done by plotting how precision, recall and accuracy improve over the number of epochs used during training of the text-toxicity algorithm.

External Server

For integration of the text-toxicity and speech-to-text algorithms into the intervention system in Unity, an external server was set up on the PythonAnywhere website. The text-toxicity analysis model could not be directly implemented into Unity due to Unity's incompatibility with looping, which is found in the bidirectional layer of the model. The server on PythonAnywhere hosts the text-toxicity algorithm and allows for a communication protocol with the intervention system in Unity. JSON (JavaScript Object Notation) data is exchanged between Unity and the server. Speech audio is inputted into Unity, converted to text using the Speech-to-Text API in Unity, sent to the text-toxicity algorithm hosted in the server, and the text-toxicity scores can be sent back to Unity.

4.1.2 Audio-Toxicity Analysis

The goal for the development of the audio-toxicity analysis algorithm was to have a program that takes datasets of audio clips labelled with their respective emotions, trains a machine learning model using those pairings, and uses that model to measure the audio-toxicity of any inputted audio clip from outside of the datasets that were used.

The research that is available involves studies on Speech Emotion Recognition (SER), but typically for more general applications such as online forums. SER is used in this thesis and involves the process of recognizing human emotion by processing audio signals, often using NLP methods for feature extraction and classification [5]. To accomplish this goal, the Speech Emotion Recognition tutorial in the form of a Jupyter notebook by Shivam Burnwall was referenced [25]. The datasets that were used include the TESS, SAVEE, RAVDESS and CREMA-D datasets [26, 27, 28, 29] composed of a combined 12162 categorized speech audio clips from a mixture of actors (both male and female) saying various phrases. The audio clips were categorized using emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). These datasets were used because they offer a large amount of varied data with a diverse set of speakers and several emotions.

Pre-processing of the audio clips involved data augmentation and feature extraction. Data augmentation is a technique used to create more varied data by taking the current set and applying perturbations to the audio in the form of noise injection, time shifting, and pitch altering. Augmenting the data in this way, can allow for the model to become invariant to small perturbations in the audio. An example of the augmentation of a waveform from one audio clip is shown in Figure 4.5 below.

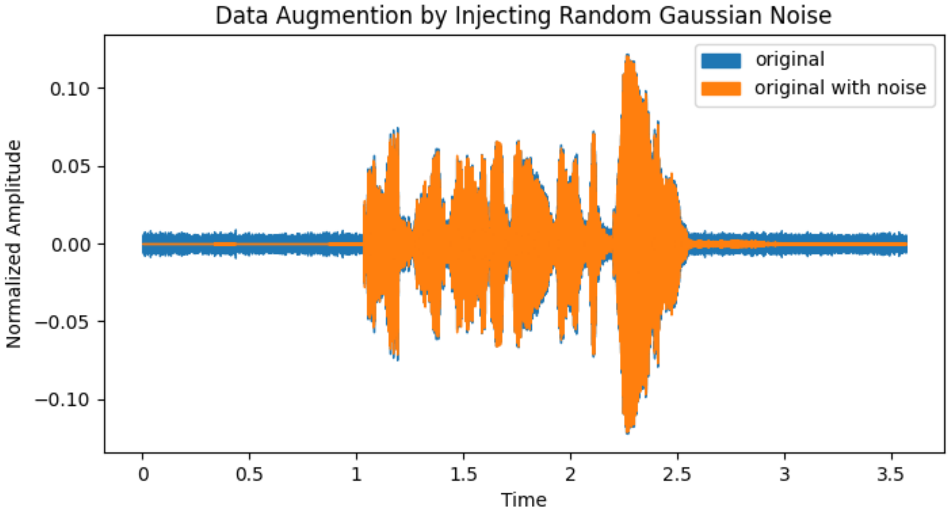


Figure 4.5: An example of one waveform that was augmented through injection of noise. The inner orange waveform represents the original, and the larger blue waveform represents the original injected with noise.

Audio feature extraction involves collecting quantifiable data of the features in the audio since unique patterns can correspond to certain emotions. Since audio-toxicity analysis algorithms are under researched, this step involved more experimentation to determine which audio features are measurable and inform toxicity. Some higher-level features of speech are shown in Figure 4.6. The audio features that are commonly used are often found by conducting calculations from the spectrograms and waveplots of audio. Some sample plots for three emotions are shown in Figures 4.7 and 4.8. These figures display clear differences, but also similarities between emotions.

To determine which audio features to use for the audio toxicity model, the book Music Data Mining by Li et al. [30] was referenced to gain a general idea of which features may be most fruitful for the task. In Chapter 5, “Mood and Emotional Classification”, Mitsunori Ogihara and Youngmoo Kim report on previous attempts that have been made to classify mood of music tracks through acoustic data analysis. They state that the highest correct classification of 61.5% on the MIREX mood data was done by Tzanetakis in 2007 using the audio features MFCC, spectral shape, centroid, and rolloff features through an SVM classifier [30]. Although an SVM classifier was not used for this intervention system, and it takes speech instead of music, this example poses a strong reference for audio feature selection for mood and emotion. A hill-climbing algorithm was made to determine which combination of features allow for the

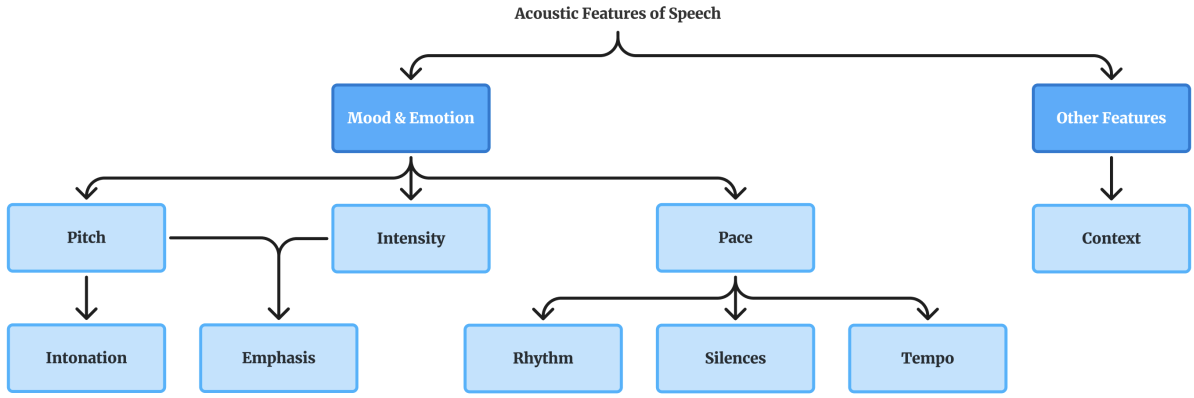


Figure 4.6: High-level acoustic features of speech and their interconnections [30, 17].

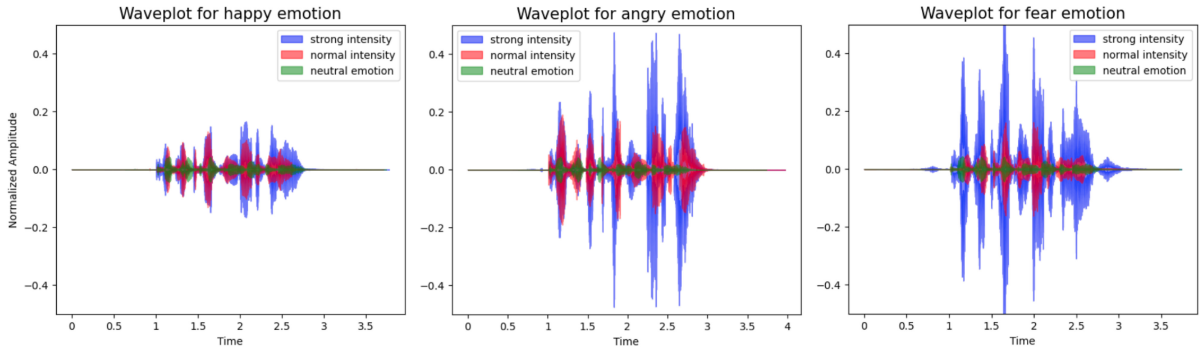


Figure 4.7: Sample waveplots for happy, angry and fear emotions.

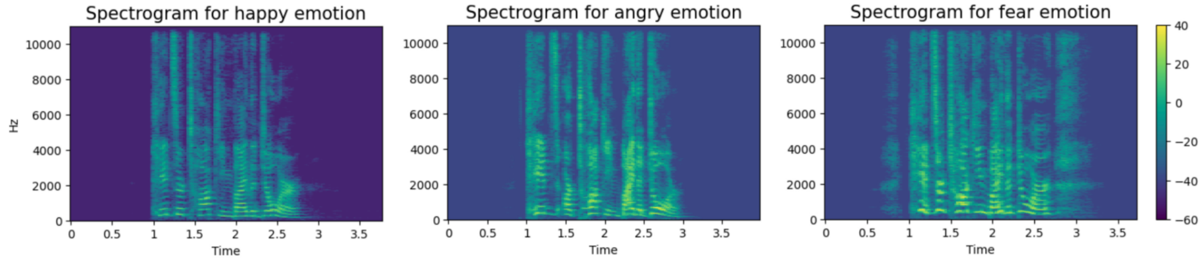


Figure 4.8: Sample spectrograms for happy, angry and fear emotions.

greatest precision and accuracy. The features listed by Tzanetakis were indeed found to be most effective, and additional features such as spectral flux, melspectrogram, root mean square of energy, entropy of energy and zero crossing rate were also tested. The optimal audio features were found to be Tzanetakis' listed features as well as entropy of energy, melspectrogram and zero crossing rate, because they resulted in a maximum accuracy of 70.09%. The Python code using the Librosa library to calculate these audio features is shown below.

```python
# MFCC
mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)

# MelSpectogram
mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)

# Spectral Flux
spectral_flux = np.mean(librosa.onset.onset_strength(y=data, sr=sample_rate).T,
    axis=0)

# Spectral Centroid
spectral_centroid = np.mean(librosa.feature.spectral_centroid(y=data, sr=
    sample_rate).T, axis=0)

# Spectral Rolloff
spectral_rolloff = np.mean(librosa.feature.spectral_rolloff(y=data, sr=
    sample_rate).T, axis=0)

# Spectral Spread
spectral_spread = np.mean(librosa.feature.spectral_bandwidth(y=data, sr=
    sample_rate).T, axis=0)

# Entropy of Energy
frame_energies = librosa.feature.rms(y=data, frame_length=2048, hop_length=512)
    [0]
probabilities = frame_energies / np.sum(frame_energies + 1e-10)
entropy_of_energy = -np.sum(probabilities * np.log2(probabilities + 1e-10))

# Zero Crossing Rate
zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
```

Once the data was pre-processed using data augmentation and feature extraction, it was passed into TensorFlow's Sequential API again, formed by layers of TensorFlow's Conv1D, MaxPooling, Dropout, Flatten and Dense layers as per the recommendation of the tutorial. However, there are limitations to this model, such as the fact that toxicity-specific audio data is sparse, and the datasets that were used consisted of generic phrases meant for day-to-day conversation.

Overarching Model

An attempt was made by Yang to create a combined model that uses both text and audio toxicity analysis in conjunction by inputting the scores of both models to get a stronger measure of toxicity. Due to scope limitations, this model was not completed but may be revisited in future work.

4.2 Design of the Interface & Interaction

The following section outlines the process of using user experience principles to design the interventions for the system, the interface elements shown in-game that act as a form of communication with the player, and the external interface elements used to present the background processes for exhibition of the prototype. The three major iterations of the design process include the wireframed prototype, the live three-dimensional game prototype pre-testing, and the final refined prototype post-testing. The final refined prototype for this thesis was not tested again and represents a work-in-progress framework at this stage.

4.2.1 Player Experience Considerations for Interventions

A goal of this intervention system is to optimize for player experience and cooperation. For the intervention system to fulfill these goals, the responses to toxicity need to be designed with carefully considered variables. Some variables that can be considered are visuals, interactivity, fairness, and accuracy of the system response. High interactivity and visuals can act as more of a distraction and punishment for toxic language, which can feel unfair to players if the system is not perfectly accurate. The more accurate the system is in detecting toxicity, the more likely that it can feel fair for players. The relationships of these variables are described by the proportionality statement below:

$$cooperation \propto experience \propto \frac{(accuracy)(fairness)(positivity)}{(visuals)(interactivity)} \tag{4.1}$$

In following this statement, visuals were kept to a minimum, and a small level of interactivity was integrated. However, as a player engages in more toxic behaviors, more visuals would appear through the system's responses and more interactivity would be required for understanding and redeeming their behavior. Success would ultimately be measured through the experience of the system from the perspectives of the players and the developers, and through the players' use of the intervention system's features.

4.2.2 Iteration 1: Wireframed Prototype

A flag system was used to indicate the different levels of warning and penalty statuses that a toxic player may be given. Flag systems are widely used from sports to legal and medical applications. Due to its widespread use, flag systems are familiar and intuitive for many. In this case, the yellow flag represents a warning, and the red flag represents a penalty.

Aggressors receive three yellow flags that act as warnings, with the opportunity to send a peace offering of some form that fits the context of the game. Peace offerings allow aggressors to redeem their behaviors and go back down one yellow flag. After a fourth instance of toxicity, aggressors receive a red flag with the penalty of being muted or suspended. For instances where players are wrongly accused by the system, a claim can be submitted to a human validator with a copy of the voice recording.

The number of yellow flags that should be given before a red flag would differ based on the game and the amount of moderation that the developer seeks. For the convenience of testing the prototype, players receive up to three yellow warning flags, and then ultimately a red flag (Figure 4.9).

Designing Interface Elements for In-Game Responses

To display these features to the players in-game, the following banners in Figures 4.10 and 4.11 were designed to consume minimal screen space (low visuals), with language written in a positive tone (high positivity). A reporting pane (Figure 4.12) was also designed for players to quickly (low interactivity) report players or view their own status during gameplay.

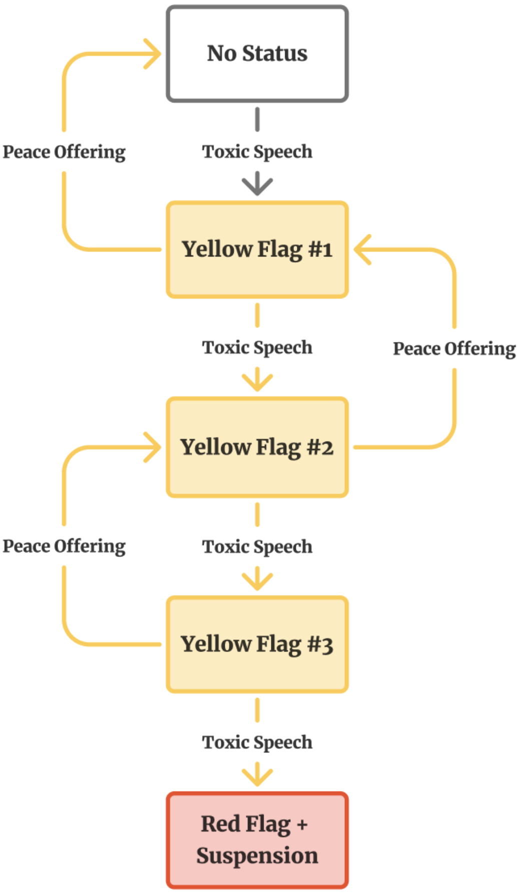
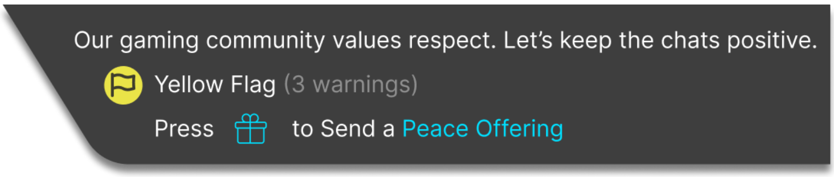


Figure 4.9: The flow of the flag system used in the intervention system for the first iteration. There were 3 yellow flags for warning, one final red flag for a suspension penalty, and peace offering options for redemption of toxic behavior.

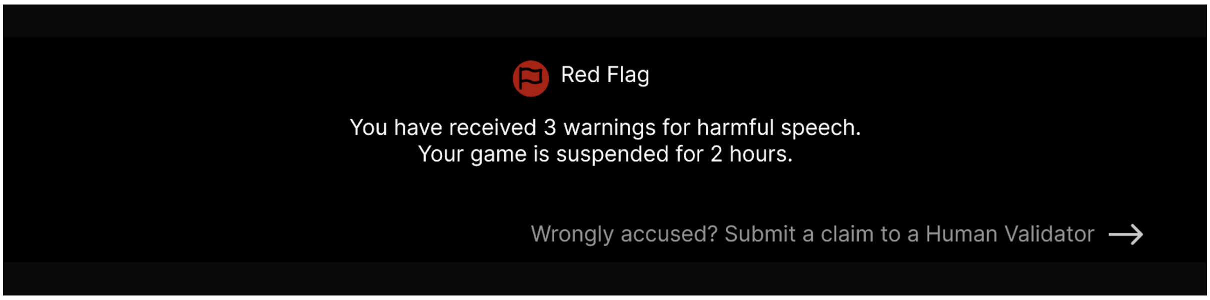


Our gaming community values respect. Let's keep the chats positive.

Yellow Flag (3 warnings)

Press 🎁 to Send a Peace Offering

Figure 4.10: Pop-up banner for yellow flag warning of Iteration 1.

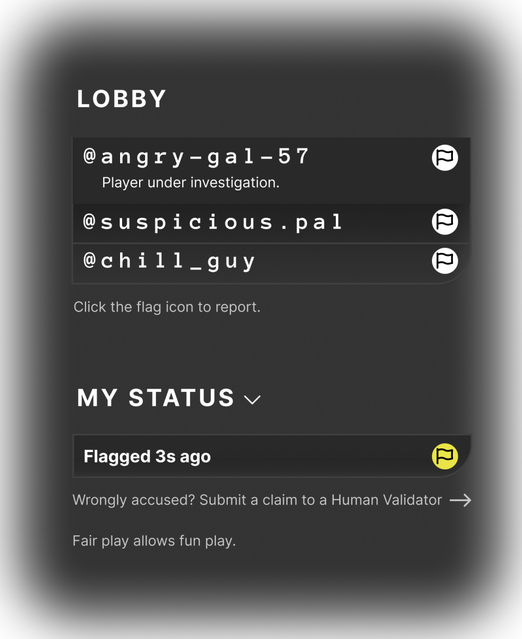


Red Flag

You have received 3 warnings for harmful speech.
Your game is suspended for 2 hours.

Wrongly accused? Submit a claim to a Human Validator →

Figure 4.11: Pop-up banner for red flag and suspension of Iteration 1.



LOBBY

@angry-gal-57

Player under investigation.

@suspicious.pal

@chill_guy

Click the flag icon to report.

MY STATUS

Flagged 3s ago

Wrongly accused? Submit a claim to a Human Validator →

Fair play allows fun play.

Figure 4.12: Reporting pane for reporting other players and viewing personal toxicity status of Iteration 1.

3D Game Environment & Display of Backend Processing

For the presentation of the intervention system prototype, a simple game environment was created to simulate the use of the intervention system in a real game, and the background processes were shown through statistics and toxicity scores on the side. An early wireframe of this is provided in Figure 4.13, consisting of buttons for recording speech, a display area for the transcribed text with visual cues to indicates states (inactive, recording, etc.), an area to display the toxicity scores, and a game wireframe.

An earlier version of the prototype was made using the wireframe in Figure 4.13, and simply involved a single stand-alone game wireframe, with the intervention system responses overlaid on top (Figure 4.14). This prototype was used during a demo to faculty and students in the Digital Futures program. The demo initiated many thought-provoking questions and discussions about the ethical dilemmas and scenarios that moderation entails. Examples include issues such as privacy and consent, voice isolation techniques, target game genres and more. These discussions incited further review of related literature to answer these questions, which provided insight to the design of later iterations.

4.2.3 Iteration 2: Live 3D Game Prototype

To create a more realistic and interactive simulation of how the intervention system would be used during real gameplay, the game wireframe was upgraded to a simplified three-dimensional first-person shooter game. Before developing the game in Unity, the user experience technique of storyboarding was used to visualize the main player flows that this game would involve.

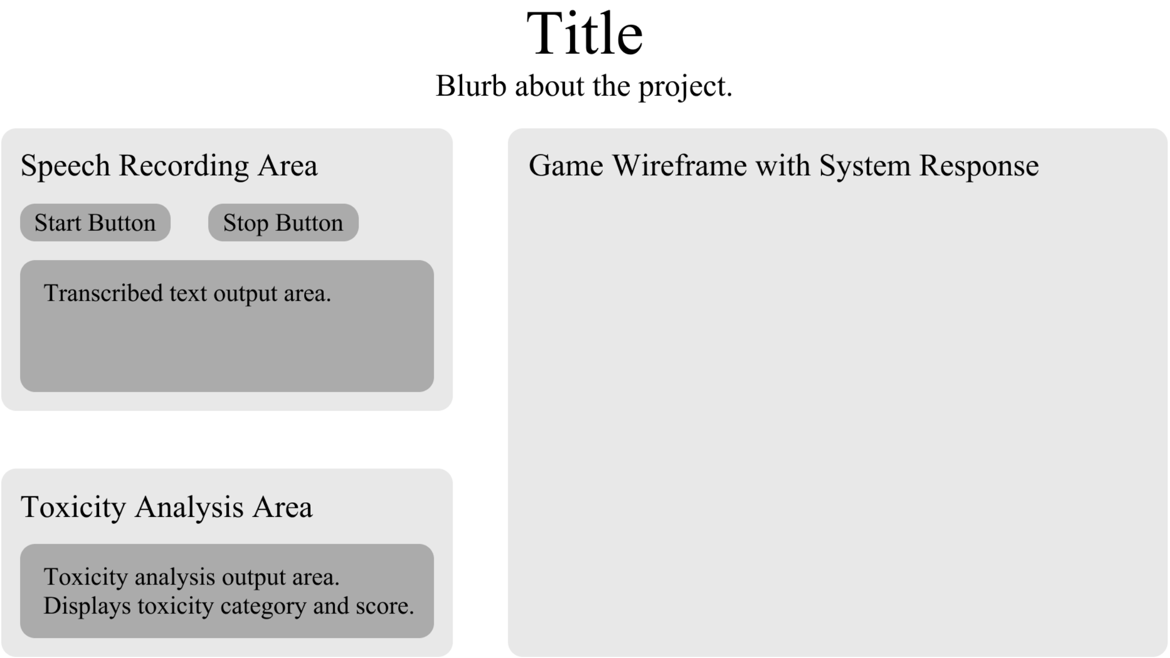
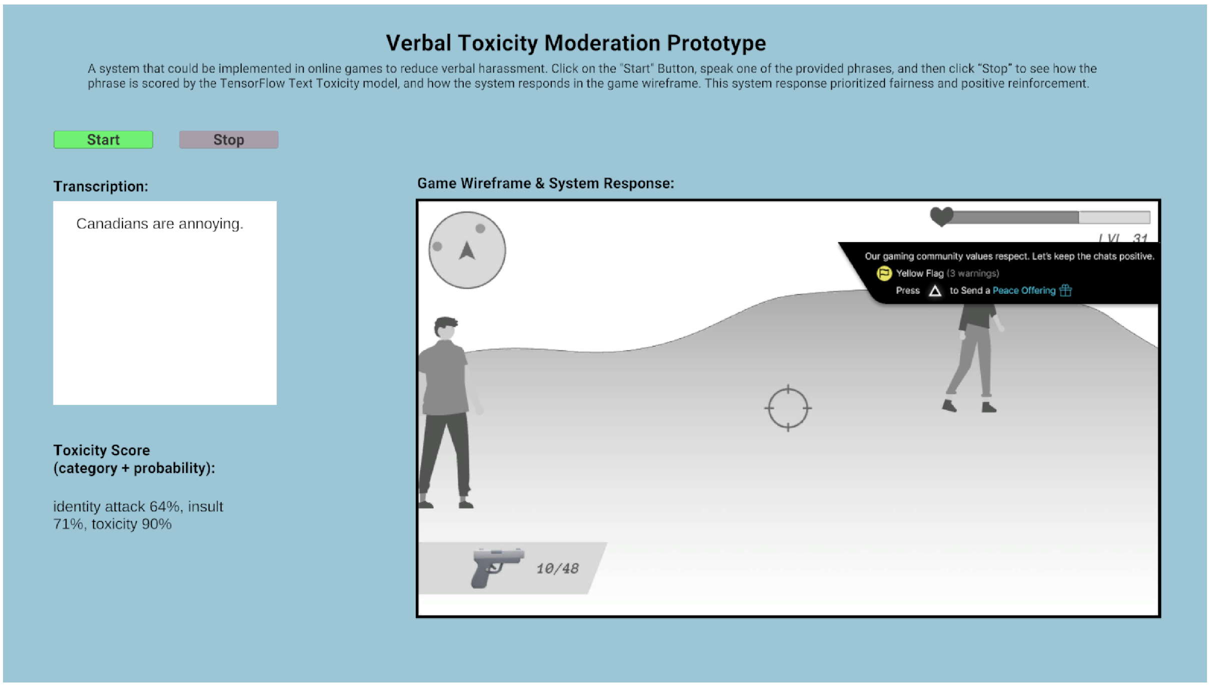

Title
Blurb about the project.
Speech Recording Area
Start Button
Stop Button
Transcribed text output area.
Toxicity Analysis Area
Toxicity analysis output area.
Displays toxicity category and score.
Game Wireframe with System Response

Figure 4.13: Wireframe of interface for Iteration 1.



Verbal Toxicity Moderation Prototype

A system that could be implemented in online games to reduce verbal harassment. Click on the "Start" Button, speak one of the provided phrases, and then click "Stop" to see how the phrase is scored by the TensorFlow Text Toxicity model, and how the system responds in the game wireframe. This system response prioritized fairness and positive reinforcement.

Transcription:

Canadians are annoying.

Toxicity Score (category + probability):

identity attack 64%, insult 71%, toxicity 90%

Game Wireframe & System Response:

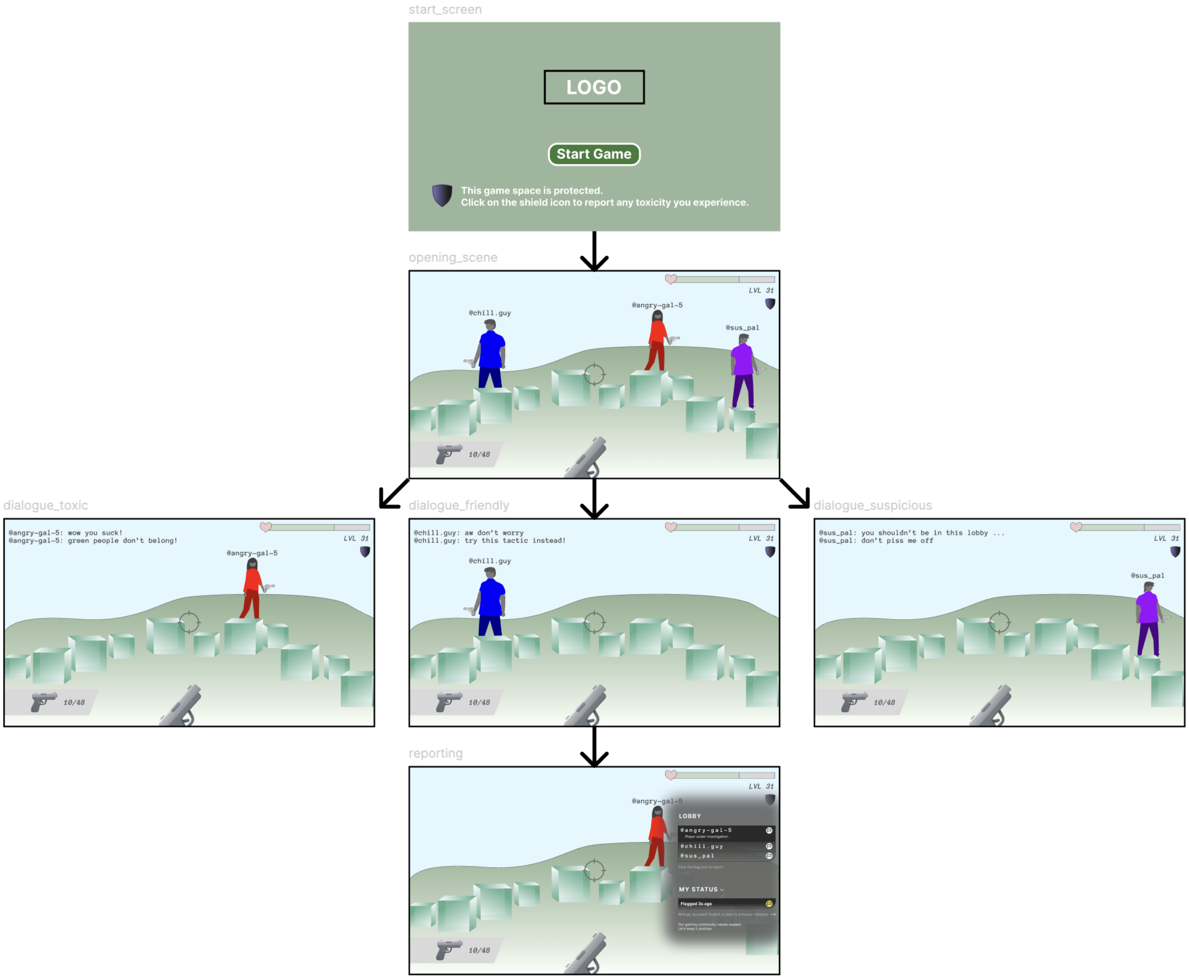
Our gaming community values respect. Let's keep the chats positive.

Yellow Flag (3 warnings)

Press △ to Send a Peace Offering

Figure 4.14: Iteration 1 full prototype made in Unity, using the wireframe in Figure 4.13.

Figure 4.15 shows one flow that was used to visualize the process of a player entering the game, interacting with the scripted characters in the game and opening the reporting pane. Three scripted non-playable characters with three different personalities were used: an angry toxic player, a calm non-toxic player, and a questionable player. These three personalities were used for player testing participants to test how the intervention system works for different levels of toxicity. The character dialogue was scripted to ensure that the toxicity content used in the dialogue was controlled since the use of a generative artificial intelligence tool for dialogue could be unpredictable and insensitive.



This game space is protected.
Click on the shield icon to report any toxicity you experience.

@angry-gal-5: wow you suck!
@angry-gal-5: green people don't belong!

@chill.guy: aw don't worry
@chill.guy: try this tactic instead!

@sus_pal: you shouldn't be in this lobby ...
@sus_pal: don't piss me off

@angry-gal-5
Player under investigation.
@chill.guy
@sus_pal
Click the flag icon to report.

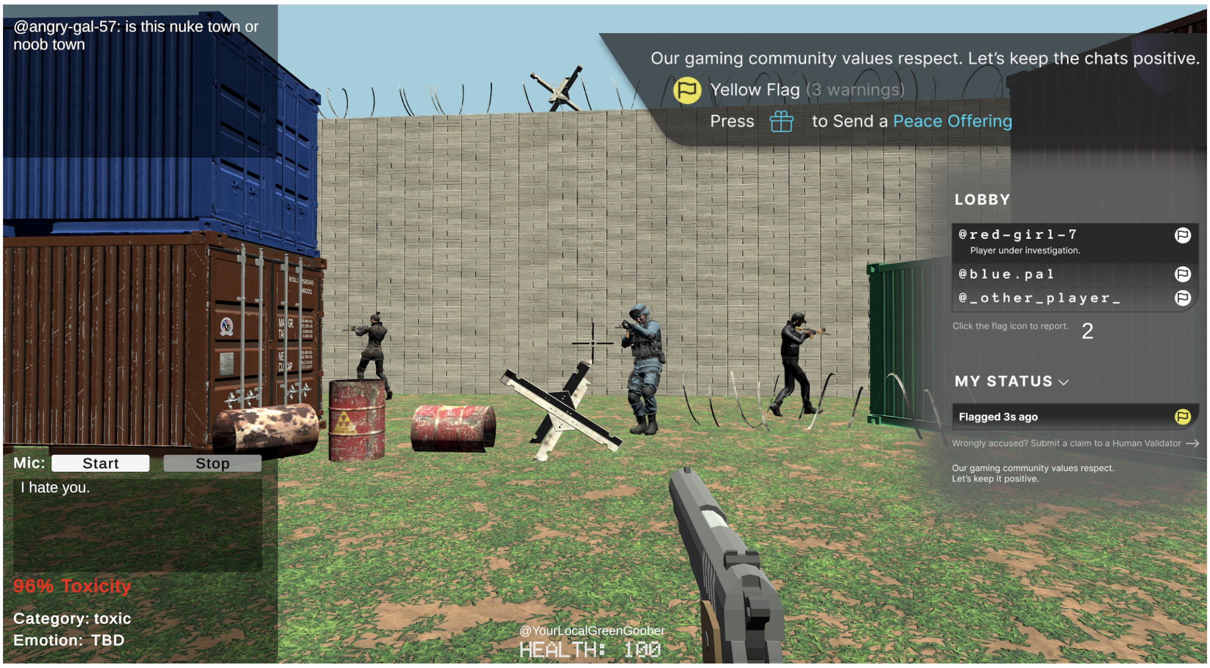
Flagged 3s ago
Wrongly accused? Submit a claim to a Human Validator →
Our gaming community values respect.
Let's keep it positive.

Figure 4.15: Snippet of storyboarding for the player flows in the 3D game of Iteration 2.

To develop the 3D FPS game in Unity, a YouTube tutorial series was referenced [31]. The displays for the background processes that were shown in Figures 4.13 and 4.14 were instead overlaid onto the 3D gameplay itself to simplify the interface for users. A screenshot of the updated prototype is shown in Figure 4.16 below. This prototype was used for player testing.

4.2.4 Player Research

To test the live 3D game prototype and to collect more insight from a diverse set of players, player surveys and player testing were conducted. The surveys were administered in the form of questionnaires that ask about demographics, their playstyles, and responses to outlined toxicity scenarios. The questionnaires provided insight to each participant's background in relation to gaming and toxicity, giving context to their responses during player testing. The player testing involved having the participants experiment with the different features of the prototype, such as the speech toxicity detection, flag system, reporting pane, and interaction with the characters. There were 10 total participants, and the conditions for their participation were that they are adults, at least moderately active online multiplayer game players, and have either witnessed or experienced toxicity.



@angry-gal-57: is this nuke town or noob town

Our gaming community values respect. Let's keep the chats positive.

Yellow Flag (3 warnings)

Press to Send a Peace Offering

I hate you.

96% Toxicity

Category: toxic

Emotion: TBD

@red-girl-7

Player under investigation.

@blue.pal

@_other_player_

Click the flag icon to report.

Flagged 3s ago

Wrongly accused? Submit a claim to a Human Validator →

Our gaming community values respect. Let's keep it positive.

Figure 4.16: Screenshot of the 3D game and interventions from Iteration 2.

Wijkstra et al.'s [7] methodology used in their systematic literature review of intervention systems was referenced to establish a procedure of conducting the surveys and testing. The following preliminary questions were established prior to completing the activities. These questions indicate which areas were focused on, and particularly refer back to the criteria of creating a system that optimizes player experience, player-system cooperation, and correction of toxic behavior:

1. Do the players feel satisfied with the responses that the system has to toxicity, specifically in terms of the flag system and the reporting system?
2. Does the intervention system flow seamlessly with their gameplay? Does it impact their experience and immersion in the gameplay?
3. Does the system encourage the players to cooperate with it in combatting toxicity?

Questionnaire

The questionnaire first requested demographic information to understand what kind of communities the participants might identify with, and to ensure that the voices of a diverse set of players were being included. Most of the participants were from the age groups of 18-22 and 23-27 years old, with 1 participant in the 33-37 age group. There was representation from the target identity groups (outlined in [1]) of African American, Woman, Muslim, Asian American, Latinx, LGBTQ+, Disability Status, and Jewish players.

Second, the questionnaire also attempted to gauge the type of player they are. The participants were asked to select from Bateman and Boon's model of seven player archetypes and the Five Factor Model of Behaviors, as cited by Caci and Dhou [14]. Participants were informed that they could select more than one option from these models. There were some indications of which participants tended towards deviant or non-deviant behaviors in later questions, but there were no strong correlations found with the player archetypes and behaviors. However, the small quantity of participants posed a limitation to determine this. The only significant trend was that 90% of participants selected the "Seekers are driven by interest and curiosity" and "Achievers are motivated by long-term achievements" player archetypes, as well as the "Openness regards the tendency to be informed, creative, insightful, curious, and to have a variety of experiences" behavior. This trend can simply point to the type of players that may have been overrepresented in the player research.

Third, participants were also asked for short answer responses on how they would react and feel in three different toxic scenarios. In the first scenario, the participant is a witness, in the second scenario, they are the victim of verbal toxicity, and in the third scenario, they are the victim of action-based toxicity. The 3 scenarios are provided below:

- Scenario 1: You are on the same team as your female friend. During gameplay, some male players in the match start making inappropriate comments about her voice and skill level. These comments escalate into unsolicited sexual remarks in the game chat.
- Scenario 2: During an online multiplayer match, a player starts using racial slurs directed at you upon assumption of your ethnic background (whether correct or not) simply because you made a minor mistake. You become the subject of verbal abuse through voice chat, which disrupts gameplay and creates a hostile environment.
- Scenario 3: In a 2 versus 2 player team-based strategy game, your teammate intentionally sabotages your own team by repeatedly destroying your structures, refusing to help during critical moments, and purposely feeding the opposing team simply because they find it funny. They do this not because they dislike the game, but to ruin the experience for you out of personal amusement.

In response to Scenario 1, all participants stated that they would feel negative emotions such as disgust, disappointment, and discomfort. There was also a common theme of lack of surprise with sexism towards female players. Seven out of ten participants indicated that they would respond to the aggressor in defense of the female player. An interesting outcome was that three of these seven participants mentioned the use of a response that would potentially get flagged for toxicity in this system. One female participant stated that she would not respond out of personal safety, and the remaining participants indicated that they would engage in action-based responses such as team killing. A notable result was that only one participant mentioned reporting, which is indicative that players are less likely to report as witnesses.

For Scenario 2, 90% of participants mentioned only defensive behaviors; they would not speak back, but would rather mute and either leave the game, continue the game, block the aggressor, or report the aggressor. Most players also indicated that they would try to ignore the aggressor and show no direct reaction. This result shows that players are more likely to report when they are the primary victim of the harassment, and that muting is an important function for them to have. Similarly, for Scenario 3, 70% of players stated that they would take defensive and avoidant behaviors such as leaving, reporting, or staying silent. The remaining

3 players would take more confrontational responses such as letting the aggressor know about their feelings and reprimanding their behavior. The results from Scenario 3 further confirm the results from Scenario 2 regarding greater chance of action from victims than witnesses.

When asked how comfortable they are with toxicity, 60% of players selected "I feel neutral. I'm used to it", and 30% chose "I get a little bothered, but not enough to leave", demonstrating they either feel uncomfortable in a toxic environment or they are used to it. Furthermore, participants were also given space to indicate if they are looking for any specific changes in moderation of toxicity. Multiple participants favored speech warnings, participation suspensions, and muting through real-time artificial intelligence. A couple of participants suggested the implementation of a system that rewards non-toxic players, and punishes toxic players through prompt, effective and neutral support to help both parties. The former two suggestions align well with this project, but the latter suggestion would need to be explored in future work.

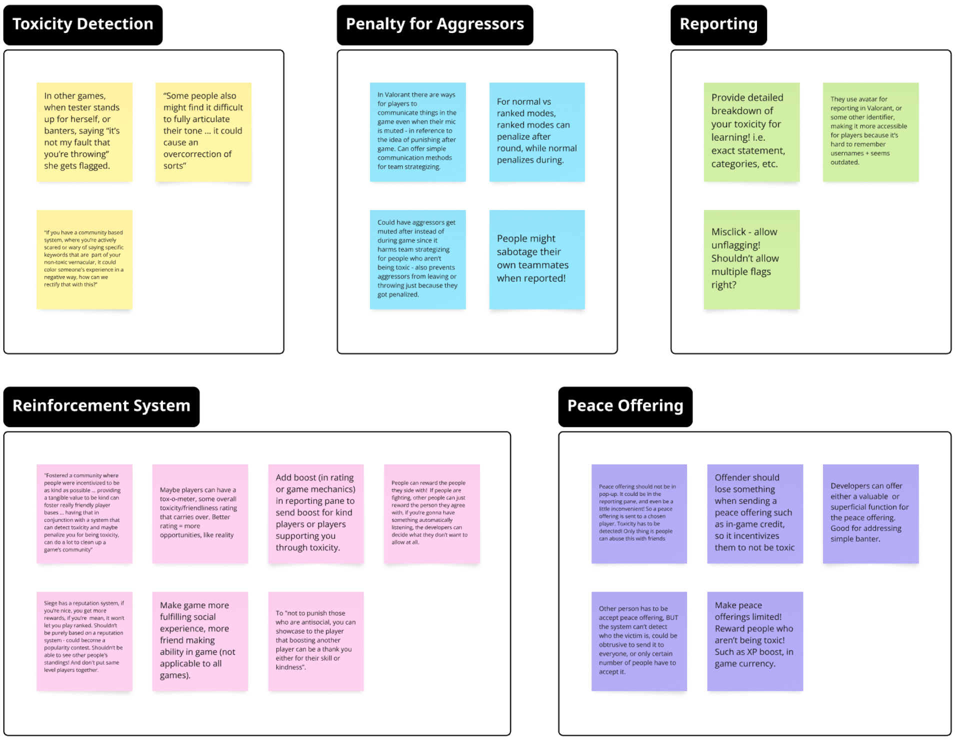
One participant also mentioned a need for more action-based toxicity moderation for griefing and throwing, as well as loss mitigation of ranked points to support victims. Ranked points simply refer to points that players can gain to move up the skill-based ranks [32], often found in competitive games. Loss mitigation is not yet a standardized term in the gaming community but generally refers to a system put in place to reduce the number of ranked points that players lose following a loss when there are forces out of their control that may have contributed to the loss. For instance, as indicated by Riot Games' Support Page for their game League of Legends [6], they award "consolation league points" when a player loses a match in which one of their teammates, that is not in their pre-made group, has been reported of going away-from-keyboard, throwing the match, or leaving mid-way [32]. The player testing participant's mention of ranked points loss mitigation for toxicity is a novel idea and would be very important for providing a form of consolation to players that may be victim to action-based toxicity. Since action-based toxicity is not in the scope of this thesis, this will be considered for future work.

Overall Questionnaire & Player Testing Results

After completing both activities, the results were synthesized by conducting thematic analysis through affinity diagramming (Figure 4.17). By creating sticky notes of the main suggestions and quotes made by participants, grouping them based on similarity, and then creating labels for the groups, five categories were formed. The categories are described below, in a problem-solution format. Minor usability and bug fixes are not reported in this paper due to the lack of meaningful relevance and were directly applied to the system.

Problem 1: Toxicity Detection Dilemmas - When players want to respond to aggressors in a defensive manner, they tend to lean into toxic speech as well, which can get them flagged. To resolve this issue, players need to be encouraged to defend themselves in a non-toxic manner. A solution that was applied for this was using more messaging and feedback to assure players that they have support from the system in flagging their aggressors. Otherwise, the decision was made that out of fairness, victims may still get flagged if they use toxic speech, even out of defense.

A few participants were also worried that the system might overcorrect for toxicity through use of a confrontational tone. It is also important to ensure that the toxicity detection algorithms



Toxicity Detection

In other games, when tester stands up for herself, or banters, saying "it's not my fault that you're throwing" she gets flagged.

"Some people also might find it difficult to fully articulate their tone … it could cause an overcorrection of sorts"

"If you have a community based system, where you're actively scared or wary of saying specific keywords that are part of your non-toxic vernacular, it could color someone's experience in a negative way, how can we rectify that with this?"

Penalty for Aggressors

In Valorant there are ways for players to communicate things in the game even when their mic is muted - in reference to the idea of punishing after game. Can offer simple communication methods for team strategizing.

For normal vs ranked modes, ranked modes can penalize after round, while normal penalizes during.

Could have aggressors get muted after instead of during game since it harms team strategizing for people who aren't being toxic - also prevents aggressors from leaving or throwing just because they got penalized.

People might sabotage their own teammates when reported!

Reporting

Provide detailed breakdown of your toxicity for learning! i.e. exact statement, categories, etc.

They use avatar for reporting in Valorant, or some other identifier, making it more accessible for players because it's hard to remember usernames + seems outdated.

Misclick - allow unflagging! Shouldn't allow multiple flags right?

Reinforcement System

"Fostered a community where people were incentivized to be as kind as possible … providing a tangible value to be kind can foster really friendly player bases … having that in conjunction with a system that can detect toxicity and maybe penalize you for being toxicity, can do a lot to clean up a game's community"

Maybe players can have a tox-o-meter, some overall toxicity/friendliness rating that carries over. Better rating = more opportunities, like reality

Add boost (in rating or game mechanics) in reporting pane to send boost for kind players or players supporting you through toxicity.

People can reward the people they side with! If people are fighting, other people can just reward the person they agree with, if you're gonna have something automatically listening, the developers can decide what they don't want to allow at all.

Siege has a reputation system, if you're nice, you get more rewards, if you're mean, it won't let you play ranked. Shouldn't be purely based on a reputation system - could become a popularity contest. Shouldn't be able to see other people's standings! And don't put same level players together.

Make game more fulfilling social experience, more friend making ability in game (not applicable to all games).

To "not to punish those who are antisocial, you can showcase to the player that boosting another player can be a thank you either for their skill or kindness".

Peace Offering

Peace offering should not be in pop-up. It could be in the reporting pane, and even be a little inconvenient! So a peace offering is sent to a chosen player. Toxicity has to be detected! Only thing is people can abuse this with friends

Offender should lose something when sending a peace offering such as in-game credit, so it incentivizes them to not be toxic

Developers can offer either a valuable or superficial function for the peace offering. Good for addressing simple banter.

Other person has to be accept peace offering, BUT the system can't detect who the victim is, could be obtrusive to send it to everyone, or only certain number of people have to accept it.

Make peace offerings limited! Reward people who aren't being toxic! Such as XP boost, in game currency.

Figure 4.17: Affinity diagramming for thematic analysis of questionnaire and player testing results.

do not automatically flag a sentence simply because they have sensitive words that are typically used in toxic speech. Certain sensitive words, even the ones that are considered slurs, can sometimes be used by members of that community in peaceful conversation. The threshold for toxicity should be kept relatively high to avoid having too many false positives but should ultimately be in the discretion of the game developer.

Problem 2: Aggressor Penalties - A common finding across player testing sessions and questionnaires was a widespread agreement on muting as a resolution. However, one participant mentioned that if aggressors are muted or suspended during gameplay, it can interfere with team strategizing, and muted players may resort to griefing or throwing (intentionally losing the game) as a means of revenge. This can impact the ability for their teammates to succeed and impede on their experience since winning is an important aspect of competitive or ranked games. A potential solution for this is to allow muted individuals to still communicate team strategy via alternate methods such as pinging and to only suspend players after a match is completed. This is an idea that was also supported by said participant. For repeated offenders that face suspension, the developers can make the choice of when aggressors are removed. An idea suggested by a participant that often partakes in competitive games is that for ranked modes, aggressors can get suspended after the round, and for normal modes they can get suspended during.

Problem 3: Reporting Issues - Victims might accidentally report the wrong player, and in the second iteration of the prototype, players can only report once. To avoid this issue, more identifiers of other players should be provided in the reporting pane when possible, such as player display pictures, avatars, roles, and more. Players should also be able to unclick the reporting button for some time to undo their mistake. They should be allowed to submit another report, up to a maximum of 3 reports per game, a minimum of 30 seconds after their previous report. These timings were discussed with participants that voiced this issue. Players can also be provided tags of key areas of toxicity to choose from when they submit a report as well. These tags can provide more information to the system and allow the players to describe their reports in a convenient manner. Furthermore, these tags can be presented to the aggressors so that they can receive clarity regarding their reports and potentially reflect on them. Some potential tags include "identity hate", "threat", "obscenity", "other", and more. An idea was suggested to also include tags for action-based toxicity such as griefing, however this idea was not implemented since there is currently no way for the system to verify this and can lead to players taking advantage of the tag. To detect griefing, developers would need unique checks in place for detection. This topic may be covered in future studies.

Another common discussion point was regarding the types of penalties that would be appropriate for toxic speech. There was a consensus among participants that muting is the optimal solution for aggressors. There were frequent references to Apple's password security feature that Apple refers to as "escalating time delays" [33]. To discourage brute-force password attacks, increased time delays are applied between password attempts following incorrect password inputs [33]. For instance, after four incorrect password attempts, users get locked out of their devices for 1 minute. After another incorrect attempt, the lockout duration increases to 5 minutes, and then 15 minutes, and so on [33]. The use of escalating time delays is a common tactic in cybersecurity practices, but it does not have a standardized name. Apple's feature was referenced by participants due to widespread familiarity, and connections were made to progressively increase the penalty for repeated offenders. This idea was implemented by establishing a similar system in which the first red flag that aggressors receive comes with microphone muting for five minutes. If the aggressor receives more red flags following their initial muting penalty, the duration gets longer, until after a certain number of red flags, the player can get suspended from participation in certain gameplay. The exact durations of these penalties would require further user research to finalize.

Problem 4: Peace Offerings - A frequent discussion point with participants was the question of what should be offered through peace offerings without giving players the opportunity to take advantage of the function. It was agreed upon that aggressors should have to send the peace offering through the reporting pane instead of the yellow flag pop-up, since it should not be very convenient as a means to disincentivize players from abusing the peace offering. There should also be a limit of three peace offerings, and only superficial rewards such as commendations and statuses should be offered instead of functional ones such as gameplay boosts. Following many discussions with participants and lab members regarding whether the peace offerings should be superficial or functional, superficial offerings were found to be the better option. Functional offerings are more likely to be taken advantage of and can shift the focus from peace resolution to trade of in-game goods. It was also considered whether the aggressor should lose what the victim gains through a peace offering to avoid a trade of goods, but that was decided against since it would decrease the incentive for aggressors to use peace offerings and the use of superficial rewards addresses this anyway.

Problem 5: Insufficient Positive Reinforcement - Participants suggested that players should be incentivized to be kind through some tangible value. They shared support with the idea of having “ranking” systems for behavior as found in some games such as Tom Clancy’s Rainbow Six Siege [5]. Participants suggested that the reporting pane should also allow players to boost the ratings of friendly players, which would carry over rounds or sessions of gameplay and can only be viewed by the player. A better rating can be achieved by being friendly and receiving boosts from other players, or by not losing the default friendliness rating through toxic behavior. Better ratings can provide more opportunities, whereas players with very low ratings may be restricted from certain multiplayer modes. This kind of system can also aid friend-making in game by placing players that give each others boosts in the same lobbies for future rounds.

4.2.5 Iteration 3: Final Refined Prototype

These changes were applied to the prototype for use in the final exhibition. Another round of testing would be required to further refine the prototype and make it more suitable for implementation in games.

Some code and links to the videos of the final prototype can be found in the author’s GitHub page: https://github.com/jessica-patel/Toxicity_Research

4.2.6 Exhibition

The intervention system prototype was exhibited at the DFX Show located at the OCAD University Waterfront Campus. The exhibition welcomed many guests of varying backgrounds, experience levels and perspectives. Consequently, there was a larger and more diverse group of exhibition participants than player testing participants. An observation that was made was the range of ways that players may exhibit toxic behavior. There were toxic comments spoken into the prototype during the exhibition that would not be detected as toxic since they are not present in the datasets that were used to train the NLP algorithms. This result encourages the need for a larger number of diverse datasets for training of detection algorithms to ensure that a majority of toxic comments are flagged.

5 Discussion

The final prototype of the intervention system accomplishes the goals of detecting toxicity through backend NLP algorithms, moderating toxicity using appropriate interface elements and working to prevent verbal toxicity with critical interventions. In reference to the optimization conditions of the final prototype, player-system cooperation was fulfilled by using assuring messaging from the system to show its support in handling toxic incidents, as well as making reporting functions easily available to victims. Player experience was accounted for through the use of user experience heuristics, and by ensuring that the intervention system does not disrupt or infringe upon gameplay. Finally, toxic behavior correction was prioritized using all four methods of operant conditioning throughout the intervention system.

A recurring theme throughout this research that was frequently voiced during player testing is that *it depends on the game.* As previously discussed in Section 2.1.3 of the Literature & Contextual Review, When and Where – Genres and Games, the toxicity levels in a gaming environment heavily depends on the genre or type of game, and the game itself. User research participants often repeated these words when questioned about certain intervention methods, indicating uncertainty in a "one-size-fits-all" intervention system. This research has reinforced that the intervention system would need to be adaptable to each game, and that the prototype made in this project can act as a preliminary form of an intervention system framework.

There were also challenges and limitations in this research. There were no major correlations found between player archetypes and toxic habits. Larger sample sizes would be required for better analysis. Rather it was found that most players tend to identify themselves with many archetypes and behavior types, indicating that it is difficult to categorize in general and that players are very diverse. Furthermore, the audio-toxicity analysis algorithm can still use further improvement since the datasets that were used only had tags for emotion, when instead they should be tagged for mood or toxicity itself. Emotion alone is not a sufficient measure for toxicity, which involves more aspects, such as tone.

Given the complexity of toxicity and the diversity of players, making design decisions for the intervention system was sometimes very difficult. Despite spending significant time and effort, and sharing countless conversations with players and academics, certain areas such as the applications of positive and negative reinforcement would require further refinement. A common challenge that arose while designing the interventions and a common topic of conversation during player testing was the following question: how should the interventions and interfaces be designed to decrease the chances of players taking advantage or dodging the attempts at moderation and correction of toxic behavior? Refinements were made to the prototype to prevent this from happening, such as rethinking the form of reward used in peace offerings, however this is an ongoing process. Due to these challenges and limitations, the current state of the prototype acts as a work-in-progress to eventually create an intervention system framework that can be tailored and implemented into various games.

6 Conclusions & Future Work

The text toxicity and audio toxicity analysis algorithms resulted in accuracies of 81% and 70%, respectively, which fulfilled the purposes of this intervention system. These results confirm the effectiveness of audio features related to spectral shape and energy in the classification of emotion. However, the lack of game-specific and toxicity-specific data poses a limitation in the use of these detection algorithms in gaming spaces. An intervention system was designed using interface elements and interventions that introduce novel and iterated forms of toxic behavior correction, player-system cooperation and player experience. This was done using a flag system and a pane that can be opened during gameplay to report and reward fellow players. The design decisions and implementations were made in accordance with the chosen heuristics from Hochleitner et al.'s (2015) heuristic evaluation framework for games [8].

For future work, the limitations would need to be addressed to achieve more accurate results in detection and greater external validity to larger populations of gamers. Alternate modalities of toxicity were unaddressed in this thesis such as action-based toxicity, which can occur in the same environments in which verbal toxicity takes place. However, the moderation of action-based toxicity would require more individualized focus and review. Furthermore, the use of context in toxicity detection was not considered in this project but is an emerging topic of interest in moderation research [17]. Context is a major factor that humans use to understand the true meaning, and toxicity levels of something that is said. Consideration of context in the future may permit yet more accuracy in the detection, and ultimately, the elimination of toxicity.

Bibliography

[1] Anti-Defamation League, “Hate is no game: Hate and harassment in online games 2023,” tech. rep., Anti-Defamation League, New York, NY, USA, Feb. 2024.

[2] O. Lopez-Fernandez, A. J. Williams, M. D. Griffiths, and D. J. Kuss, “Female gaming, gaming addiction, and the role of women within gaming culture: A narrative literature review,” *Frontiers in Psychiatry*, vol. 10, 2019.

[3] Anti-Defamation League and Fair Play Alliance, “Disruption and harms in online gaming framework,” tech. rep., Center for Technology and Society, Dec. 2020.

[4] E. D. Liddy, “Natural language processing,” in *Encyclopedia of Library and Information Science*, New York: Marcel Decker, Inc., 2nd ed., 2001.

[5] S. Ghosh, S. Lepcha, S. Sakshi, R. R. Shah, and S. Umesh, “Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances,” *arXiv preprint arXiv:2110.07592*, 2021.

[6] J. Cai and D. Y. Wohn, “Categorizing live streaming moderation tools: An analysis of twitch,” *International Journal of Interactive Communication Systems and Technologies*, vol. 9, no. 2, pp. 36–50, 2019.

[7] M. Wijkstra, K. Rogers, R. L. Mandryk, R. C. Veltkamp, and J. Frommel, “Help, my game is toxic! first insights from a systematic literature review on intervention systems for toxic behaviors in online video games,” in *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion ’23)*, (New York, NY, USA), pp. 3–9, Association for Computing Machinery, Oct. 2023.

[8] C. Hochleitner, W. Hochleitner, C. Graf, and M. Tscheligi, “A heuristic framework for evaluating user experience in games,” in *Game User Experience Evaluation* (R. Bernhaupt, ed.), pp. 187–206, Springer, 2015.

[9] T. M. Leeder, “Behaviorism, skinner, and operant conditioning: Considerations for sport coaching practice,” *Strategies*, vol. 35, no. 3, pp. 27–32, 2022.

[10] M. Wijkstra, “Fighting toxicity through positive and preventative intervention,” in *Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY ’24)*, Oct. 2024.

[11] G. Ambrose and P. Harris, *Basics Design 08: Design Thinking.* AVA Publishing, 2010.

[12] K. L. Gray, *Race, Gender, and Deviance in Xbox Live: Theoretical Perspectives from Virtual Margins.* Waltham, MA, USA: Anderson Publishing, 1st ed., 2014.

[13] K. M. Christopherson, “The positive and negative implications of anonymity in internet social interactions: ’on the internet, nobody knows you’re a dog’,” *Computers in Human Behavior*, vol. 23, pp. 3038–3056, Nov. 2007.

[14] B. Caci and Z. Dhou, “The interplay between artificial intelligence and users’ personalities: A new scenario for human-computer interaction in gaming,” in *HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games*, pp. 619–630, Springer, Oct. 2020.

[15] R. Bartle, “Hearts, clubs, diamonds, spades: Players who suit muds,” *Journal of MUD Research*, vol. 1, no. 1, p. 19, 1996.

[16] M. Salter, “From geek masculinity to gamergate: The technological rationality of online abuse,” *Crime, Media, Culture*, vol. 14, pp. 247–264, Aug. 2018.

[17] Z. Yang, Y. Maricar, M. Davari, N. Grenon-Godbout, and R. Rabbany, “Toxbuster: In-game chat toxicity buster with bert,” *arXiv preprint arXiv:2305.12542*, pp. 1–11, 2023.

[18] Z. Yang, N. Grenon-Godbout, and R. Rabbany, “Game on, hate off: A study of toxicity in online multiplayer environments,” *ACM Games*, pp. 1–12, Jun. 2024.

[19] Modulate, “Toxmod for gaming,” 2025. Accessed: Mar. 14, 2025.

[20] Hasso Plattner Institute of Design at Stanford, *An Introduction to Design Thinking Process Guide*, 2010.

[21] N. Renotte, “Commenttoxicity,” 2022. Accessed: Sep 14, 2024.

[22] Jigsaw and Google, “Toxic comment classification challenge,” 2017.

[23] D. Fesalbon, “Gosu ai english dota 2 game chats,” 2023. Accessed: Mar. 19, 2025.

[24] A. Jain, “Cyberbullying dataset,” 2023. Accessed: Mar. 19, 2025.

[25] S. Burnwal, “Speech emotion recognition,” May 2020.

[26] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (tess),” 2010. Accessed: Mar. 19, 2025.

[27] S. Haag and P. Jackson, “Surrey audio-visual expressed emotion (savee),” 2009. Accessed: Mar. 19, 2025.

[28] S. R. Livingstone and F. A. Russo, “Ryerson audio-visual database of emotional speech and song (ravdess),” 2018. Accessed: Mar. 19, 2025.

[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crowd-sourced emotional multimodal actors dataset (crema-d),” 2015. Accessed: Mar. 19, 2025.

[30] T. Li, M. Ogihara, and G. Tzanetakis, eds., *Music Data Mining.* Boca Raton, FL: CRC Press, 2011.

[31] GameDev.tv, “Fps full game tutorial — unity — part 1 - basic movement,” March 2021. Accessed: Mar. 19, 2025.

[32] R. Games, “Mmr, rank, and lp,” 2024. Accessed: Mar. 14, 2025.

[33] A. Inc., “Apple platform security,” 2024. Accessed: Mar. 14, 2025.

Ludography

[L1] Rockstar North, “Grand Theft Auto V,” 2013. Platforms: PlayStation 3, Xbox 360, PlayStation 4, Xbox One, Windows, PlayStation 5, Xbox Series X/S.

[L2] Infinity Ward, “Call of Duty,” 2003. Platforms: Microsoft Windows, Mac OS X, N-Gage, PlayStation 3, Xbox 360.

[L3] Valve Corporation, “Dota 2,” 2013. Platforms: Microsoft Windows, Linux, Mac OS X.

[L4] Bungie, “Destiny 2,” 2017. Platforms: PlayStation 4, Xbox One, Windows, Stadia, PlayStation 5, Xbox Series X/S.

[L5] Ubisoft Montreal, “Tom Clancy’s Rainbow Six Siege,” 2015. Accessed: Mar. 14, 2025.

[L6] Riot Games, “League of Legends,” 2009. Accessed: Mar. 14, 2025.

[L7] Mojang Studios, “Minecraft,” 2011. Accessed: Mar. 14, 2025.