# Pixel to Platform: Reforming Online Toxic Communities

CS4991 Capstone Report, 2023

Dominic DaCosta
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
dld5mxn@virginia.edu

## ABSTRACT

Microsoft Corporation, a software company headquartered in Redmond, WA, is a leader in the technological industry, including being a leader in the video game creation space, being the owner of one of the one of the world's largest gaming systems, the Xbox. As part of this venture into the gaming world, Microsoft and Xbox own several gaming studios, all serving to produce best in class first party video game experiences for the platform. Since the turn of the century, gaming as a medium has garnered worldwide usage, now becoming one of the world's most influential mediums for all ages. As it currently stands in multiplayer live service games, players who exhibit toxic behavior are often met with a ban from the game service, ranging from a temporary hold on the account to a permanent ban. This method is ineffective at addressing the root cause of the issue, oftentimes with high recidivism from the same players. My project proposes a warning to players, restating the expected behavior and showing as a physical reminder to abide by such laws . To address this issue, as an intern within the Minecraft Player Safety team, I built an end-to-end warning system using Xbox Live Services and in-game infrastructure with the intention of changing disruptive players behavior before a ban was issued to the player. To implement this solution, I used Xbox Live services to access player metadata to issue a scaled warning, using a tiered system. The project was announced and met with positive feedback internally, and will be expected to release in the next quarterly cycle.

## 1. INTRODUCTION

Since the turn of the century, the video game industry has undergone a massive shift in consumer opinion, transitioning from a niche hobby to a one of the biggest entertainment mediums in the world, rivaling even Hollywood for consumer share. Microsoft, through its acquisition of various gaming studios including Mojang Studios - the creators of Minecraft, the highest selling video game of all time, has been at the forefront of this transformation. As proprietary owners of arguably the most influential piece of video game hardware, that being the Xbox, in addition to owning thousands of other video game properties, the focus of my project was set against this backdrop, addressing an increasingly disturbing and persistent challenge in the gaming community - toxic behavior.

Traditional methods of managing disruptive behavior, that primarily being through either permanent or temporary bans and account suspensions, have shown limited effectiveness. They often fail to address the underlying causes of such behavior, leading to high rates of recidivism. Recognizing this gap, my project during my 3 month summer

internship with the Player Safety team within Minecraft, aimed to implement a system that would not only penalize but also educate and guide players towards more positive interactions in their respective online communities.

My specific work focused on creating a phased system within the current architecture of the software. Based on extensive research conducted by other members of the team, we determined that we would utilize our integrated chat detection system, Community Sift, to first determine whether what is being said in real time is deemed to be appropriate or inappropriate, based on a multitude of categories, ranked by least to most egregious. If deemed to be not within our guidelines, the idea was to have this detection tool trigger automatic warnings, elevating in severity as they continuously ignore our warnings. These progressive stages in how we handle this behavior gives players with good intentions or those who are not aware of current guidelines a chance to learn from their mistakes and no longer continue their behavior. In theory, this would lead to lower recidivism rates amongst toxic players in our playerbase. In addition, I also worked on a frontend moderation tool for internal use for our moderators and engineers to be able to trigger this system manually if necessary, as well as

## 2. RELATED WORKS

The initiative to develop a warning system for managing disruptive behavior in multiplayer live service games has been an ongoing initiative by some of the top minds in the world. To deal with this problem, extensive research in behavioral psychology and game design are required. One piece of related work in this field is B.J. Fogg's concept of Captology, specifically found in his work "Persuasive technology: using computers to change what we think and do" ,which delves into computers as "persuasive technologies" (Fogg, 2002). This framework is pivotal in understanding how online environments, like gaming platforms, can be designed to positively influence user behavior. Another significant contribution to this area is the work of Kwak et al. (2015), who, in their study presented at the ACM International Conference on Human-Computer Interaction, explored the complexities of toxic behavior in online gaming communities. Their findings particularly highlights the necessity for context-sensitive intervention strategies, focusing on the intricate nature of player interactions within these virtual environments.

In addition to these studies, the importance of implementing behavioral interventions in online settings is demonstrated in the research by Smith and Christakis (2008), published in the Journal of Social and Clinical Psychology. Their research provides evidence that well-designed online interventions can effectively modify social behavior, a principle directly applicable to online gaming.

## 3. PROJECT DESIGN

In initial discussions into the design of this project with my team and other engineers related to areas of the codebase I would be working with, we ultimately landed on a strategic approach when selecting the technology stack and system design. This decision was driven with safety and quality above all else. It was imperative that this project seamlessly integrate into existing features and not unknowingly break anything. This was a more difficult task than it sounds, as the entirety of the codebase is reliant on other portions. The first piece to consider was the integration with Xbox Live Services. This integration was crucial for accessing and utilizing player metadata, enabling the

issuance of tailored, context-specific warnings. This was a crucial aspect, as other portions of this warning system relied on grabbing player metadata to understand how frequently this player has been offending our policies.

The second piece to consider when designing this project was utilizing our existing Community Sift tool for nuanced chat detection. The incorporation of Community Sift's advanced chat detection capabilities allowed for real-time monitoring of player interactions. The nature of the tool, which is able to detect chat and able to understand a multitude of languages and each of their nuanced meanings, allowing for a higher percentage of chat abuse to be detected. This tool was pivotal in identifying instances of toxic behavior, serving as the backbone of the warning system's operational framework.

Another thing we had to take into account was the existing infrastructure. The seamless embedding of the warning system within Minecraft's existing in-game infrastructure ensured that the player experience remained uninterrupted. This aspect of the design was critical in maintaining the immersive nature of the game while introducing a new layer of player interaction and safety.

Beyond the coding aspects, the project's success was equally dependent on the surrounding infrastructure and tooling. This encompassed a multitude of processes. The first, being Iterative Testing and Quality Assurance. Thanks to the help of pre-existing quality assurance teams within the organization, I was in constant communication with this team to discuss possible vulnerabilities within the design. However, this also encompassed constant testing, through unit and integration testing. This ensured each component of the warning system functioned effectively within the complex ecosystem of Minecraft and met the high-quality standards set.

For integration into the codebase, github was used. One piece that was decided very early, was the languages and tools to be used to complete this. As essentially all of our backend was written in C and C++, this is what I primarily used throughout my work, as well as JavaScript for the frontend. These tools allowed for a high level of performance with faster processing times.

## 4. RESULTS

My involvement in the project during my internship led to significant contributions in both the implementation and further refinement of the warning system. I gained deep technical knowledge in the design architecture of the software the organization used, but also gained knowledge in the areas of networking, frontend development and game development.

The integration of the system with Xbox Live Services and Community Sift was a pivotal achievement. It demonstrated the system's capability in real-time behavior monitoring and intervention, marking a significant milestone in online gaming safety. Leading a demonstration of the system's capabilities to executives of Xbox and Mojang Studios, I was able to showcase its practical application, through a video demo, in identifying and addressing various levels of toxic behavior. It received overwhelmingly positive feedback, even discussing possible future iterations, such as the emerging new AI technology ChatGPT to aid in the fight against toxic behavior in online communities. This presentation not only served as a crucial validation of the system's effectiveness but also highlighted its potential impact on player safety and community standards beyond Minecraft, with discussions about possibly

expanding this into the main Xbox ecosystem.

## 5. CONCLUSION

With the completion of this internship project for the Minecraft Player Safety team, it serves as the start to an advancement in addressing the challenges of toxic behavior in online gaming communities. The development of an end-to-end warning system, integrating Xbox Live Services and Community Sift, signifies a leap forward in the realm of digital community management and player safety. This project not only bridges the gap in existing methods of managing disruptive behavior but also sets a new standard in proactive and educational interventions within gaming environments. Its success underlines the importance of innovative technological solutions in fostering safer, more inclusive online spaces. The project goes beyond mere functionality; it is a testament to the power of technology and software in reshaping online interactions and community standards, leading to healthier online communities.

## 6. FUTURE WORK

Following the completion of my presentation to Xbox and Microsoft Leadership, the feature is scheduled to be launched in Q3 of the team's quarterly schedule. The team's focus will be on polishing performance and further ensuring code quality is where it needs to be through continuous rigorous testing.

## 7. ACKNOWLEDGMENTS

All of this work would not be possible without the guidance of my direct Player Safety team within Microsoft, specifically Jason Burch and Declan Hopkins who mentored me throughout this process. Their technical insights into the architecture of the codebase was integral in the success of this project.

## REFERENCES

Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do. Morgan Kaufmann.

Kwak, H., Blackburn, J., & Han, S. (2015). Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. ACM International Conference on Human-Computer Interaction.

Smith, A., & Christakis, N. (2008). Social Networks and Health. Journal of Social and Clinical Psychology.