



Breaking the Building Blocks of Hate

A Case Study of *Minecraft* Servers

The first analysis of
hate and harassment on
Minecraft server data.

A report from the ADL
Center of Technology & Society
JUL 2022



CENTER FOR
TECHNOLOGY
& SOCIETY

Our Mission

To stop the defamation of the Jewish people and to secure justice and fair treatment to all.

About the Authors

Rachel Kowert, Ph.D is a research psychologist and the Research Director of Take This. She is a world-renowned researcher on the uses and effects of digital games, including their impact on physical, social, and psychological well-being. To learn more about her work, visit www.rkowert.com

Austin Botelho is a Machine Learning Engineer at ADL's Center for Technology and Society where he leverages data and computational techniques to reduce online hate and harassment. He holds an MSc in Social Data Science from the University of Oxford

Alex Newhouse is the Deputy Director of the Middlebury Institute's Center on Terrorism, Extremism, and Counterterrorism, where he focuses on right-wing extremism, religious fundamentalism, and militant accelerationism.



ADL (Anti-Defamation League) fights antisemitism and promotes justice for all. Join ADL to give a voice to those without one and to protect our civil rights.

About

Center for Technology & Society

Launched in 2017, ADL's Center for Technology and Society (CTS) leads the global fight against online hate and harassment. In a world riddled with antisemitism, bigotry, extremism, and disinformation, CTS acts as a fierce advocate for making digital spaces safe, respectful and equitable for all people.

CTS plays a unique role in civil society by recommending policy and product interventions to elected officials and technology companies to mitigate online hate and harassment; driving advocacy efforts to hold platforms accountable and to educate their staff on current threats and challenges; producing data-driven applied research by analysts and a network of fellows, shedding new light on the nature and impact of hate and harassment on vulnerable and marginalized communities; developing tools and products that provide much needed data measurement and analysis to track identity-based online hate and harassment; and empowering targets of harassment by responding to online incidents and working with platforms to create safer online spaces for all.

Anti-Defamation League

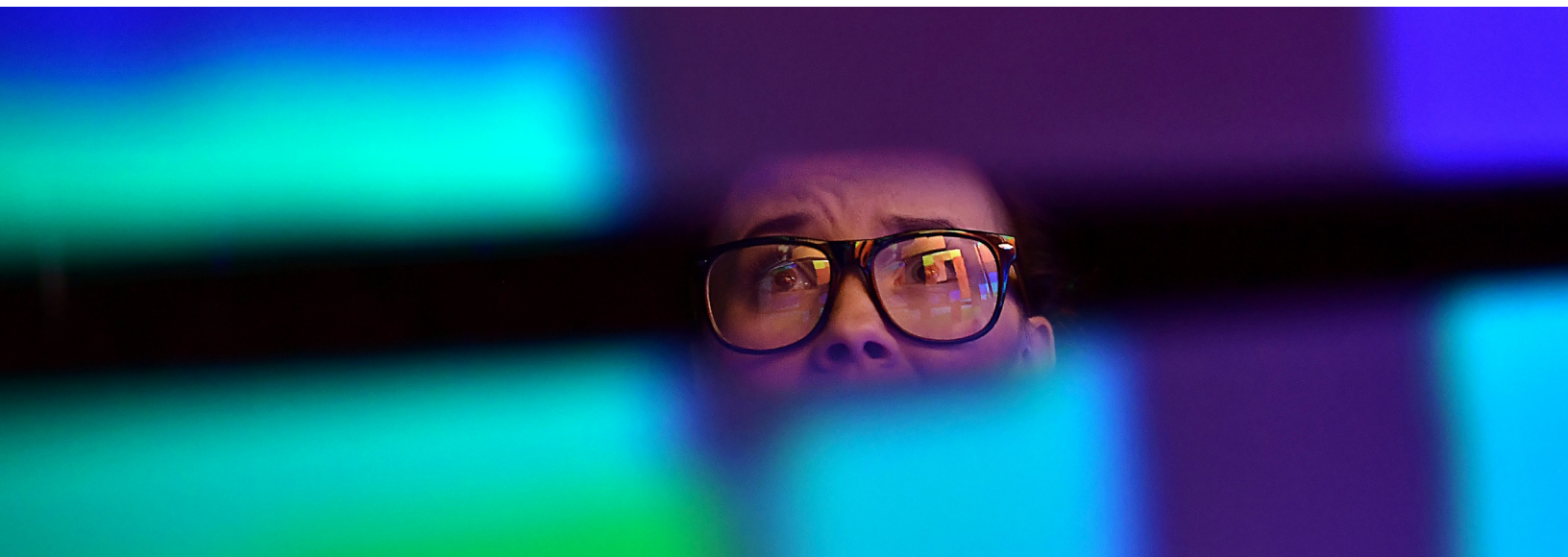
ADL is a leading anti-hate organization that was founded in 1913 in response to an escalating climate of antisemitism and bigotry. Today, ADL is the first call when acts of antisemitism occur and continues to fight all forms of hate. A global leader in exposing extremism, delivering anti-bias education and fighting hate online, ADL's ultimate goal is a world in which no group or individual suffers from bias, discrimination or hate.

Table of Contents

Executive Summary	04
Introduction	07
<i>Minecraft</i> Java Edition	09
How We Conducted our Analysis	10
Findings	
Formal reports of hate speech via GamerSafer	11
The Role of Moderation	13
Text analysis: Sexually explicit, hateful, and severely toxic interactions	14
Recommendations	16
About the Contributors	17

Executive Summary

The online game *Minecraft*, owned by Microsoft, [has amassed 141 million active users since it was launched in 2011](#). It is used in school communities, among friend groups and even has been [employed by the U.N.](#) Despite its ubiquity as an online space, little has been reported on how [hate and harassment manifest in *Minecraft*](#), as well as how it performs content moderation. To fill this research gap, Take This, ADL and the Middlebury Institute of International Studies, in collaboration with GamerSafer, analyzed hate and harassment in *Minecraft* based on anonymized data from January 1st to March 30th, 2022 consensually provided from three private *Minecraft* servers (no other data was gathered from the servers except the anonymized chat and report logs used in this study). While this analysis is not representative of how all *Minecraft* spaces function, it is a crucial step in understanding how important online gaming spaces operate, the form that hate takes in these spaces, and whether content moderation can mitigate hate.



After examining chats and user reports from data provided by GamerSafer, we found:

- **Many in-game offenders are repeat offenders.** Almost a fifth of offending users had multiple actions taken against them during the data collection.
- **Temporary bans proved to be an effective solution for reprimanding bad behavior.** Early evidence shows temporary bans to be more effective than muting in reducing the rate of offending behaviors by the moderated player.
- **Servers with in-depth community guidelines were associated with more positive social spaces.** Of the three servers reviewed, Server 3 had the most extensive community guidelines and the lowest frequency of sexually explicit, hateful, and severely toxic behavior between users, suggesting the positive impact of robust guidelines.
- **Server rules appear to matter more than moderation enforcement in shaping communication norms.** There was no effect of moderation events on rates of server-wide toxic behaviors over time. Nonetheless, rates of toxic behaviors still correlated with server staffing and rules.
- **The rates of harmful behavior differed depending on the message type.**
 - Hateful messages were 21% more likely for public chats than private ones.
 - Sexually explicit messages were 9% more likely for private chats than public ones.
- **Analysis further suggests hateful rhetoric has been normalized in gaming spaces.** The presence of slurs previously only affiliated with white nationalism and hate groups suggests the normalization of extreme language in gaming spaces.
 - Sexually explicit language occurred 3x as often as hateful language.

Based on the above findings, we suggest:

- **Increasing researcher access to data.** Granting researchers and watchdogs with access to data helps to identify and address the challenges of hateful, harassing, and toxic behavior in spaces meant to have a positive social impact.
- **Investing in content moderation efforts and robust community guidelines.** Active, effective human moderation and community guidelines are critical to reducing sexually explicit, hateful, and severely toxic behavior in gaming spaces as the sever with the most staff and most extensive guidelines had the fewest incidents of these kinds of behaviors.
- **Additional research into content moderation and complementary tools and techniques,** including the:
 - short and long term benefits of content moderation efforts. Moderator intervention seems to reduce behavior in the short term, but it remains unclear as to whether this holds over time. Future work should focus on determining the long-term effects of moderator intervention.
 - impact of tools and techniques that enable player accountability and go beyond moderation, such as player verification and globally blocklisting players banned for severe harms. These measures can add responsibility and minimize reoffending.

This work highlights the importance of granting third parties access to in-game data. Allowing researchers access to unfiltered data, while still protecting user data privacy, can provide unprecedented insight into the interactions between users within gaming spaces, and more broadly, users' interaction in online social spaces. Without this level of data transparency, the industry will neither be able to identify nor address the challenge of hateful, harassing, and toxic behavior in online gaming spaces.



Introduction

Three billion people worldwide play online multiplayer games; 97 million of them are American. Digital games have permeated almost every aspect of pop culture but the explosive growth of these spaces and communities has led to the rise of toxic gamer cultures that exclude people and worse, threaten their safety. [Hate and harassment are now widespread in gaming spaces](#), worrying parents, scholars, and policymakers.

The high prevalence of hate in online games has fueled concern that problematic language and behavior are becoming normalized within gaming communities, desensitizing individuals to hate speech and fostering polarization between communities.

Nearly all gamers have experienced harassment. [ADL's 2021 annual survey of hate and harassment in games](#) found that 5 out of 6 adults ages 18-45 and 3 out of 5 young people ages 13-17 experienced harassment. One of the more pernicious forms of online harassment is hate, which is broadly defined as speech or behavior that targets people based on their identity, including gender, race or ethnicity, religion, sexual orientation, gender identity, physical appearance, or disability. ADL's survey also found that 8% of adults and 10% of young people reported being exposed to white supremacist ideology in online games. These numbers may be much higher, as [recent research found that 64% of online players have experienced hate speech directly, and 83% have witnessed it happening to others.](#)

To gain a better understanding of hate speech within gaming cultures, ADL partnered with [Take This](#), the [Middlebury Institute of International Studies](#), and [GamerSafer](#) to examine speech patterns in public and private chat logs from private *Minecraft* Java servers. The decentralized, player-run nature of *Minecraft* Java edition provides a novel opportunity to assess hate and harassment in gaming spaces.



Extremists co-opt Minecraft, Call of Duty and more to spread messages of hate. buff.ly/3nY8RF9



9:02 AM · Sep 23, 2021 · Buffer

<https://twitter.com/CBR/status/1441025288675028992>

Minecraft Java Edition

Minecraft is a popular sandbox style game developed by Mojang Studios in 2011 and was subsequently acquired by Microsoft in 2014. Players can create and break apart various kinds of blocks in three-dimensional worlds. *Minecraft* is often framed as a “sandbox” because like a traditional playground sandbox, only a player’s imagination limits what they can do. There are no real goals or objectives except the ones that players set for themselves.

There are two versions of *Minecraft*: the Bedrock and Java editions. The Java edition primarily allows players to independently host and privately run game servers with no oversight from Mojang whereas Bedrock is primarily hosted by Mojang, which partners closely with the servers’ users on content moderation. In June 2022, *Minecraft* [announced](#) it was enabling players of the Java edition to report violative content to the central *Minecraft* team at Mojang Studios. Simultaneously, Mojang confirmed that this effort would only focus on what players report and exclude any proactive monitoring or moderation.

The size and scope of these servers can vary greatly. For example, an individual can host a small server open to a handful of people living in a particular neighborhood to be used as a creative after-school space. A server could be used by an [entire classroom](#) or a school to explore group learning in a digital space. It may also be a server hosted by an organization such as the United Nations to [provide young people with an interactive experience in urban planning](#). Since server maps are based on user-generated content, players can also spread hateful concepts and congregate like-minded people. [Private Minecraft servers have been found to host the creation and re-enactment of Holocaust concentration camps](#). Until recently, the administrators of Java servers (such as educators setting up a *Minecraft* server for their students) enacted all content moderation decisions for that server. With the recent advent of player reporting, the *Minecraft* team at Microsoft will also play a role in content moderation, though what that role is remains to be seen.

We hope this report provides a benchmark on the scope and form that hate takes in the online games millions play every day.

How We Conducted Our Analysis

We analyzed public and private chat data provided by GamerSafer from three *Minecraft* Java servers to measure the levels of hateful speech in gaming spaces. The three servers varied in size and moderation capacity.

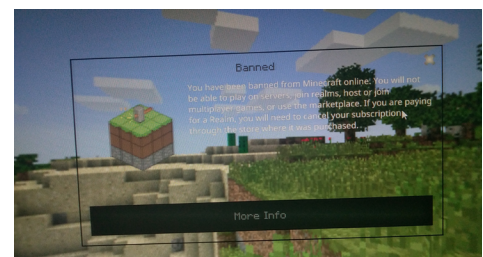
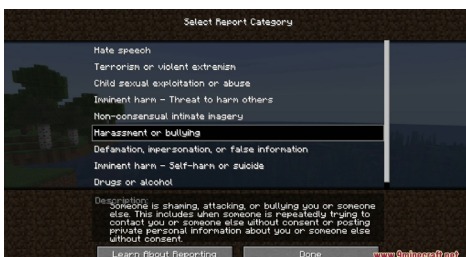
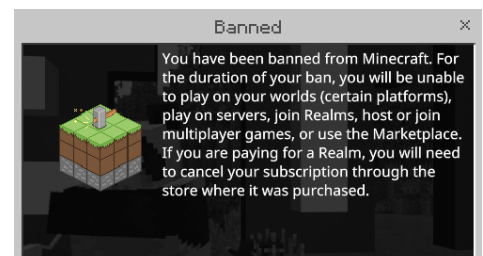
Server 1 housed a large (tens of thousands of players), primarily adolescent audience (14-18 years old), dozens of server staff, and upheld very strict rule enforcement. Competitive player versus player (PvP) play was optional on this server. Server 1 was the most active of the three servers and constituted 94% of the analyzed data.

Server 2 had a smaller (several thousand players), slightly older audience (averaging 15-20 years old) and only two staff members that implemented minimal to no rule enforcement. On this server, competitive PvP was encouraged.

Server 3 was the smallest of the three servers, housing only hundreds of players. This server was primarily an audience of older adolescents and adults (aged 16+) with a moderation team of 10 staff members. Gameplay was largely collaborative, with PvP limited to small, designated areas. Of all the servers, Server 3 had the most extensive guidelines and active moderation team but staff had trouble enforcing the rules.

GamerSafer offers a suite of safety tech solutions to gaming companies, including a special plugin for *Minecraft*, supporting server staff and moderators to manage infractions and in-game reports. Hate speech is one of [37 ban categories](#) in the *Minecraft* plugin. Tracking incidents of reported behavior among game players across servers, the report logs contain 458 total disciplinary actions against 374 unique users.

Analysis was two-fold. First, we examined the formal reports made to moderators and moderator actions in regards to any form of hateful or harassing behavior. This was done to better understand the nature of the servers and broad patterns of behavior. Secondly, we conducted a textual analysis on the in-game chat to examine speech patterns as well as identify hateful or harassing actions that may not have been logged by the moderators.



Findings

Formal reports of hate speech via GamerSafer

Many in-game offenders are repeat offenders. Almost a fifth of users (64) had multiple actions taken against them during data collection. Server 1 contained 93.9% of these reports.

Moderators have the ability to ban players. Nearly two thirds of the actions taken by moderators were bans. As seen in Figure 1, a temporary ban from the server was the primary action (46%), followed by a permanent ban (20%), and a verbal warning (19%). We do not know with certainty how closely bans were enacted following a violation.

Figure 1. Frequency of moderation intervention across the three servers

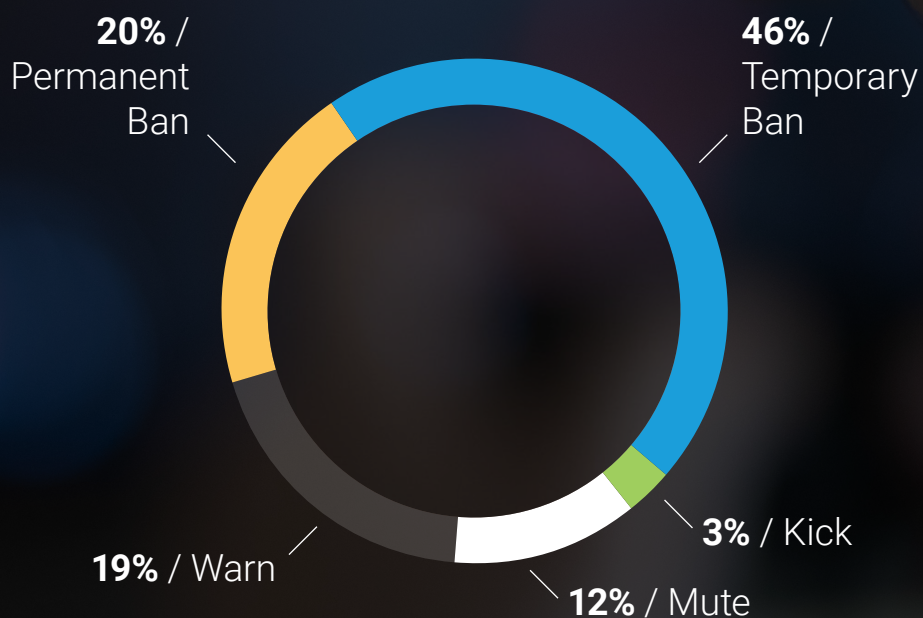
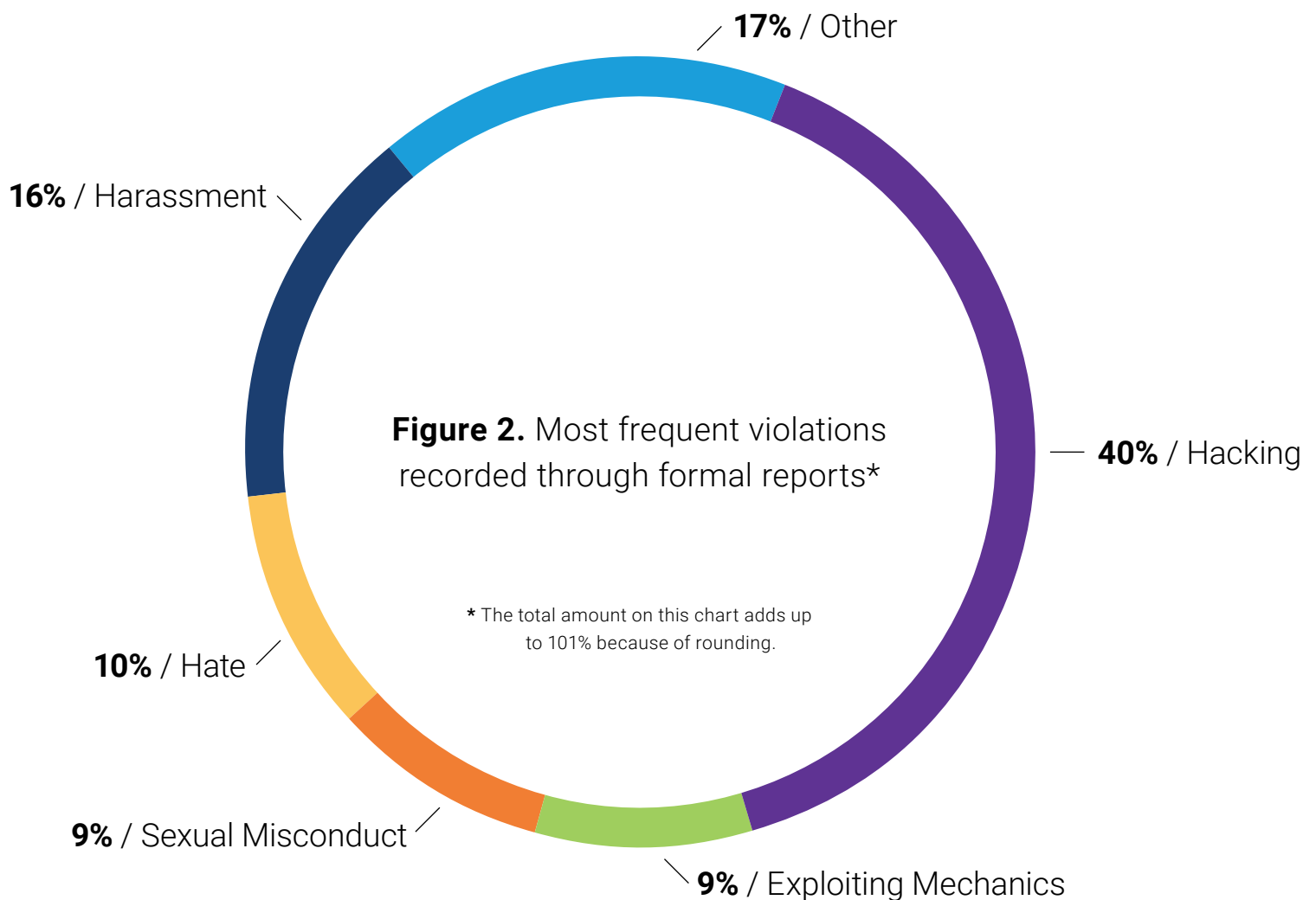


Figure 2 shows the most common reported violations reported were hacking, or trying to create an advantage beyond normal game play (40% of all reports); harassment, including hostility, spamming, bullying, trolling and threats (16% of all reports); hate, or attacks rooted in prejudice towards someone's race or ethnicity, gender, sexual orientation, and other forms of identity (10% of all reports); sexual misconduct in the form of inappropriate comments (9% of all reports); and exploiting mechanics, or using a bug or glitch in a game for an advantage (9% of all reports).





The role of moderation

Moderation efforts reduce unwanted behavior. They led to substantial reductions in sexually explicit, hateful, and severely toxic behavior across all servers. We found:

- Temporary bans were found to be the most effective for reducing the rates of offending behaviors by moderated players.
- Server 3 had the the most extensive community guidelines and the lowest frequency of sexually explicit, hateful, and severely toxic behavior between users, suggesting the positive impact of strong guidelines.
- No significant connection was found between the moderation frequency and rates of toxicity over time, suggesting moderation may be more effective as a short- rather than long-term solution. Further research is needed to understand these relationships.

Moderation, both manual and automated, is often believed as key to reducing negative behaviors in games. We assessed the relationship between moderator actions and changes in chatter behavior. To do this, we linked actions in the chat log to incidents in the report log based on their timestamp and anonymized ID. Given the report lag time, we defined incidents as any chats from the user sent within 12 hours of the report. While this excludes incidents with a longer report lag and includes benign chats where the lag was shorter, we felt this best balanced the need to link as many reports as possible without overspecifying. It is important to note that not all reports appear in the chat log as many were for in-game behaviors (i.e., non-chat related actions) or chats from affiliated Discord channels we could not access.

This method was able to link 48 incidents to the chat. We could then compare the number of sexually explicit, hateful, and severely toxic messages sent before and after the moderator action to evaluate the action's effectiveness. The data indicated that all moderation actions led to reductions in the three types of harmful behavior by the actioned user across all servers, with temporary bans being the most successful. Temporary bans reduced actioned users' sexually explicit content by 82%, hate by 93%, and severe toxicity by 85%.



Another way to evaluate the impact of moderation norms on toxic activity is to look at differences in harmful behavior across the three servers, which have different content rules and moderator staffing. We observed that the servers have differing rates for sexually explicit¹, hateful², and severely toxic content³. Our analysis shows that Server 3 had the least amount of harmful activity whereas Servers 1 and 2 were not statistically different from each other. Servers 1 and 2 were more sexually explicit (91% and 80% of all content found, respectively), severely toxic (96% and 86% of all content found, respectively), and hateful (78% and 87% of all content found, respectively) than Server 3. Notably, moderation did not seem to impact the rates of harmful behavior over time, as no significant connections were found between the moderation frequency and rates of reporting hate, explicit content, and severe toxicity. This last result may be unique to this dataset or unique to the ways in which content moderation is conducted in these servers. Further investment in research and content moderation efforts is needed to understand how moderation can best impact and reform harmful behavior in online gaming.

Text analysis: Sexually explicit, hateful, and severely toxic interactions

- There is evidence for normalization of extreme language in gaming spaces using slurs previously only affiliated with white nationalism and hate groups.
- The normalization of extreme language in gaming spaces is more common for hateful rather than sexually explicit language.

Across all three servers, we assessed 1,463,891 messages sent by 9,008 unique users from January 1st to March 30th, 2022. While hate speech was our primary incident of interest, we also analyzed the nature and prevalence of sexually explicit language (references to sexual acts, body parts, or other lewd content) and toxicity (severely hateful, aggressive, or disrespectful comments likely to make a user leave a discussion or decline to share their perspective). We wanted to provide a comparison of hate speech to other forms of disruptive online behavior.

That data was run through a series of keyword queries using Perspective⁴. Of all the logged chats, 2% (28,254 messages) could be categorized as severely toxic, 1.6% (24,462 messages) as sexually explicit, and 0.5% (7,574 messages) as hateful. Hate speech largely targeted sexuality (e.g. f*g homo) and gender (e.g. b*tch) .

1. Sexually explicit X2 (2, N = 1204658) = 237.6, p = 2.58e-52

2. Hate speech X2 (2, N = 1204658) = 62.0, p = 3.43e-14

3. Severely toxic X2 (2, N = 1204658) = 292.8, p = 2.57e-64.

4. The definition for these categories are taken from Perspective, which is an automated tool used for text analysis (<https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>). Perspective is a machine learning based tool that estimates the likelihood a piece of text contains one of its covered toxic behaviors. Like many automated tools, it can be error-prone; for example, we excluded the threat scores from our analysis as violent language often describes gameplay rather than interpersonal aggression. That said, it is widely trusted to surface toxic content at scale.

a. Public versus private chat messages

Our analysis also revealed that rates of harmful behavior differed depending on the message type. There were 21% more hateful messages in public chats versus private chats. Private chats were 9% more sexually explicit than public chats. This difference is important as it highlights the differences in levels of acceptance around hateful language in gaming spaces. It is possible that hateful and/or harassing language is more commonplace and normalized within public spaces whereas sexually explicit language is more likely to be shared user to user. It is also plausible that players use hateful language to provoke outrage or garner attention from the rest of their community. Sharing hateful language in private chats would not give the abusers the satisfaction of being seen by others. Similarly, [online radicalization in games often starts with an invitation for other members to join in or acquiesce to hateful language shared in public gaming spaces](#). Expressing this language publicly, rather than privately, identifies other members of the community who share similar sentiments.

b. Normalization of hate speech

In addition to formal reports, we examined chat logs for hate speech (e.g., gender-based slurs, insults against LGBTQ+ people) that were not formally reported by the players on the server. We discovered that such slurs and other extremist-adjacent language occurs at a detectable level and often goes unmoderated. Although these terms were used infrequently, it is noteworthy that on these heavily regulated servers, players still engage with hateful narratives.

Using keyword lists compiled by ADL and Middlebury, we searched unfiltered chat logs for language that is either overtly hateful or employs terms that originate from hateful contexts. Within the language of concern, we found that normalized language occurred most frequently—that is, hateful language integrated into colloquial conversation. For example, of all terms searched in this data set, the extreme term with the highest number of pejorative uses was “cuck.”

Cuck, short for “cuckold”, was adopted as an insult by online communities in the mid-2000s and then popularized by right-wing networks in the mid-2010s, when it was used to castigate anyone whose views were not adequately right-wing. Cuck became associated with the so-called “alt-right” movement, which [weaponized the term against political opponents to sexually or racially demean them](#). Reaching its peak relevance during the 2016 presidential election, [the term has gradually spread throughout more mainstream communities and is now used by many people who do not understand its contemporary sexist and racist connotations](#).

Users in these servers also frequently employ gendered expletives and other phrases, which is common for video game communities. Gendered curse words can contribute to intolerance toward women and LGBTQ+ people and are worth addressing. Across servers, players used “b*tch” hundreds of times and occasionally used meme misspellings like “wamen” that are [associated with the “Manosphere”](#), a loose online community that promotes masculinity, misogyny, and opposition to feminism.

Recommendations

Working with real-time chat data, this deep dive into *Minecraft* Java servers provided insight into hate and harassment within games. Drawing from our findings, we offer directions for future work within this space.

Increasing Researcher Access to Data

- This work highlights the importance of access to server-logged data for scientific analysis.
- As an industry, allowing access to unfiltered data can provide unprecedented insight into the frequency and nature of interactions between users within gaming spaces. Without this level of data transparency, the games industry cannot identify or address the challenges of hateful, harassing, and toxic behavior.
- In addition to raw data, future research should include the ability to assess the links between moderator actions and chat messages. For example, researchers must be able to determine when a moderator action is performed because of gameplay, voice chat, or text messages (in-game or on third-party software like Discord) to better understand the impact of moderation.

Additional Research Into Content Moderation and Complementary Tools and Techniques

- Moderator intervention seems to reduce behavior in the short term, but it remains unclear as to whether this holds over time. Future work should focus on determining the long-term effects of moderator intervention.
- There is an untapped opportunity to investigate the short- and long-term impact of tools and techniques that enable player accountability and go beyond moderation, such as player verification and globally blocklisting players banned for severe harms. These measures can add responsibility and minimize reoffending.

Investing in Content Moderation Efforts and Robust Community Guidelines

- Active, effective human moderation and community guidelines appear to be critical to reducing sexually explicit, hateful, and severely toxic behavior in gaming spaces as the sever with the most staff and most extensive guidelines had the fewest incidents of these kinds of behaviors.
- Industry leaders need to continue investing in moderator training to better understand and respond to toxic behaviors.

Standardize reporting categories

- To better understand the frequency and nature of hate in online spaces, we recommend an industry-wide standardization of moderation reporting, including defined categories and violating offences with clear descriptions. This would help facilitate future research, particularly in regards to documenting how moderation actions change user behavior over time. The ADL and Fair Play Alliance's [Disruption and Harms in Online Gaming Framework](#) could be used as the foundation for this effort.

About the Contributors



Take This (www.takethis.org) is a 501(c)(3) mental health non-profit that serves to destigmatize mental health challenges and provide mental health resources and support for the gaming industry and gaming communities.



ADL is the leading anti-hate organization in the world. Founded in 1913, its timeless mission is "to stop the defamation of the Jewish people and to secure justice and fair treatment to all." Today, ADL continues to fight all forms of antisemitism and bias, using innovation and partnerships to drive impact. A global leader in combating antisemitism, countering extremism and battling bigotry wherever and whenever it happens, ADL works to protect democracy and ensure a just and inclusive society for all.



Middlebury Institute of International Studies (<https://www.middlebury.edu/institute>) is a graduate school of Middlebury College, based in Monterey, California.



GamerSafer's (www.gamersafer.com) vision is to scale online safety, positive and fair play experiences to millions of players worldwide. Using computer vision and artificial intelligence, its identity management solution helps multiplayer games, esports platforms, and online communities defeat fraud, crimes, and disruptive behaviors.

Support

This work is made possible in part by the generous support of:

We are grateful to the Anti-Defamation League, Russell Sage Foundation, Carnegie Corporation of New York, and the National Science Foundation for financial support and to Samantha Luks at YouGov for outstanding survey administration assistance. We also would like to thank Virgílio A. F. Almeida, Christopher Charles, Mark Ledwich, Becca Lewis, Wagner Meira, Raphael Ottoni, Manoel Horta Ribeiro, Aaron Sankin, Robert West, and Anna Zaitsev for sharing their data with us or making it publicly available. All conclusions and errors are our own.

Anonymous (2)

The Robert Belfer Family

Dr. Georgette Bennett and Dr. Leonard Polonsky

Bumble

Crown Family Philanthropies

Electronic Arts

Grove Foundation

Walter & Elise Haas Fund

Craig Newmark Philanthropies

Quadrivium Foundation

Righteous Persons Foundation

The David Tepper Charitable Foundation, Inc.

The Harry and Jeanette Weinberg Foundation

ADL Leadership

Ben Sax

Chair, Board of Directors

Jonathan Greenblatt

CEO and National Director

Mike Sheetz

President, Anti-Defamation League Foundation

Adam Neufeld

Senior Vice President and Chief Impact Officer

Center for Technology & Society

Daniel Kelley

Director, Strategy and Operations

Austin Botelho

Machine Learning Engineer

Caroline Bermudez

Editorial Director

Take Action

Partner with ADL to fight hate in your community and beyond.

- Sign up at adl.org for our email newsletters to stay informed about events in our world and ADL's response.
- Report hate crimes and bias-related incidents in your area to your regional ADL office.
- Engage in respectful dialogue to build understanding among people with different views.
- Get involved with ADL in your region.

Featured Resources

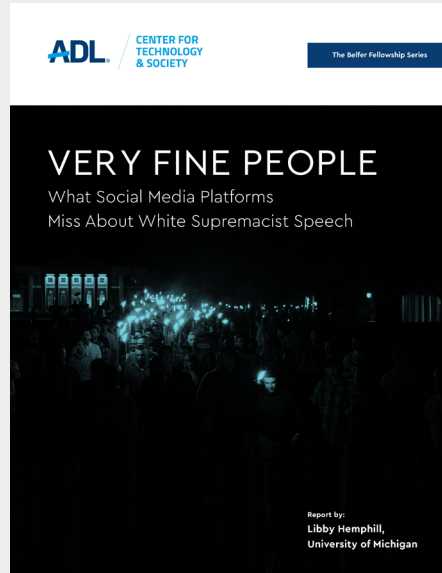
From the ADL Center for Technology & Society



Online Hate and Harassment

The American Experience 2022

<https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022>



Very Fine People

What Social Media Platforms Miss About White Supremacist Speech

<https://www.adl.org/resources/report/very-fine-people>



Hate is No Game

Harassment and Positive Social Experiences in Online Games 2021

<https://www.adl.org/resources/report/hate-no-game-harassment-and-positive-social-experiences-online-games-2021>



[adl.org](https://www.adl.org)



Anti-Defamation League



@ADL_National



@ADL_National

