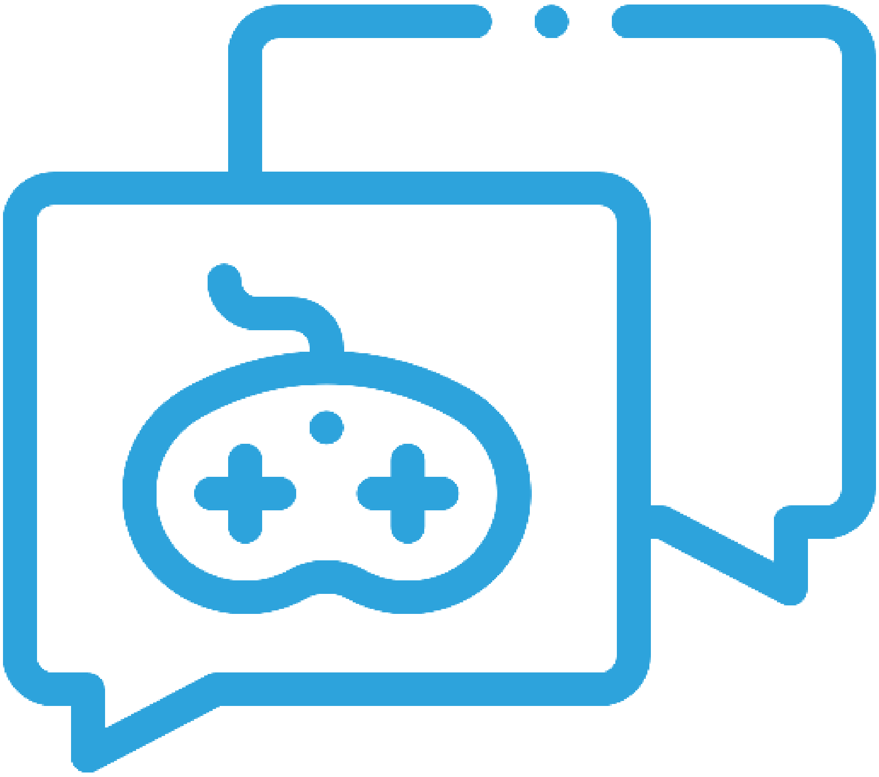
Combatting Hate & Extremism on Your Gaming Platform

The presence of cyberbullying and extremism is growing in gaming, researchers have found, while the industry's hidden metrics, insipid content moderation and head-in-the-sand attitudes get in the way of evaluating and combating the problem which affects customer safety and trust.



cleanspeak 2022

cleanspeak

WRITER
Blair Ewalt

DESIGNER
Sean Bryant

CONTACT
CleanSpeak
390 Interlocken Cres.
Suite 332
Broomfield, CO 80021

EMAIL
sales@cleanspeak.com

cleanspeak.com

A Quick Story

Misha Valencia's

12-year-old son was playing an online game with friends when a new user in the group let several others in. As The New York Times reported, those new users quickly "flooded the chat with anti-Semitic vitriol, swastikas, and neo-Nazi propaganda."

When Valencia's son tried to shut down the messages, the new users began to verbally attack him. He and his mother blocked and reported one hateful message after another but couldn't keep up with the barrage in what seemed to be a coordinated attack.

Hate speech, harassment, extremism, and targeted attacks like these have become all too common.

The problem is serious due to the fact that gaming plays a huge role in American life. Two thirds of adults and three quarters of children under 18 play video games weekly, and 74% of Americans have at least one video game player in their household. With nearly 3 billion active video game players worldwide and 227 million video game players in the U.S. alone, it probably comes as no surprise that online interactions can range from friendly and competitive to derisive and threatening with children and adults alike.

In a perfect world, gaming would be based on fairness, a sense of community, and respect. However, the truth is that sexist, racist, and other abusive hate speech and extremism in gaming can be a bullying nightmare for gaming community members and a PR nightmare for the gaming platforms themselves.

In a March 2022 study by global e-learning platform Preply, three quarters of all respondents indicated that they want speech restrictions on game platforms. In response, businesses large and small are investing in manual and automated solutions to combat offensive language, imagery, and videos in all gaming communications while simultaneously working to improve their brand image and protect user safety. Some of these solutions are proving to be highly effective; others are merely expensive and time-consuming stopgaps that fail to curb the offenses.

Here, we'll review the serious nature of hate and extremism on gaming platforms and highlight what to look for when choosing a profanity

The Growing Problem of Hate and Extremism on Gaming

B2C Companies

Tend to understand that customer satisfaction and reputation management are crucial to their success. They need to recognize the importance of monitoring the content that their platforms host to provide a safe and trusted environment for their users, comply with federal and state regulations, and manage brand perception.

Those seeking to instigate hate and violence for their ideological ends are turning toward gaming spaces in greater numbers, especially as traditional social media platforms continue to amp up enforcement of their rules and crack down on the content users share.

This problem warrants attention because gaming is big business. Recent data indicates that gaming is now the most profitable entertainment industry around the globe, generating more money each year than the movie and music industries combined. The global games market is estimated to generate $152.1 billion from 2.81 billion gamers worldwide. By comparison, the global box office industry was worth $41.7 billion in 2020, while global music revenues reached $19.1 billion.

Hate Speech & Extremism Defined

Hate Speech

Might be defined as any form of expression that expresses or incites hatred against a group or a class of persons on the basis of race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin. Yale Law School professor Robert Post notes that prohibiting or suppressing hate speech “is to forbid expression of ‘extreme’ intolerance or ‘extreme’ dislike. The qualification ‘extreme’ is prerequisite because intolerance and dislike are necessary human emotions which no legal order could pretend to abolish.”

According to the Anti-Defamation League, extremism is a “concept used to describe religious, social or political belief systems that exist substantially outside of belief systems more broadly accepted in society (i.e., ‘mainstream’ beliefs). Extreme ideologies often seek radical changes in the nature of government, religion or society.”`

Hate Speech & Extremism in Gaming

Organizations

That have platforms that rely on user-generated content are struggling to maintain customer safety and trust. Hate speech and extremism are hard to monitor and censor due to the massive quantities of text, images, videos, and content being created daily.

According to the Aforementioned Preply Study,

OVER 90%

Of gamers have experienced or witnessed emotional abuse or bullying on gaming platforms, and more than half admitted to bullying others.

NEARLY 7 IN 10

Have considered quitting due to what they've witnessed. More than two in five have experienced racism, and more than one in three have experienced hate speech on gaming platforms.

NEARLY 20%

Of gamers say they've experienced some kind of extremist content.

Offenses can include name-calling, racism, stalking, hate speech, insults, explicit language, swatting (calling emergency services with false reports of violent crimes), flaming (hostile online interactions that involve an exchange of insulting messages, or flames, between users), physical threats, and doxxing (revealing someone's sensitive information).are hard to monitor and censor due to the massive quantities of text, images, videos, and content being created daily.

The Problem is Widespread. For Example:

STEAM

The leading social platform for PC gamers, has recently gained popularity among white supremacists for being a platform, like Gab and

DISCORD

A group chatting platform for gamers, has become an online space for initiating people from the early stages of radicalization all the way to extremism.

TROLLING & HATE RAIDS

Have thrived on Twitch, an interactive livestreaming service for content spanning gaming and entertainment.

EXTREMIST

Groups have built propaganda content, including interactive Nazi concentration camps, in such video games as Roblox.

ANTI-SEMITIC,

Racist, and homophobic comments are often found on gaming platforms that allow users to stream and chat about games, such as Call of Duty and Minecraft.

Why Many Game Developers WANT to Proactively Address Toxicity

1. They don't want a game's (or their own) reputation to be associated with toxicity.

2. They want to reduce a game's "toxic" atmosphere, where players come to expect toxicity to occur anytime they play.

3. They know many gaming spaces are being exploited by extremists, and they want to make sure their community is protected.

4. They're troubled by the massive amount of vitriol on their platform, and they're ready to do something about it.

5. They want to avoid litigation, bad publicity, negative consumer sentiment – all of which can cut into the bottom line. They want to address a reputation crisis before it spirals out of control. sure their community is protected.

Why Many Game Developers DON'T Proactively Address Toxicity

1. Moderation process development and staffing can be time consuming and extremely expensive.

2. They want to keep chats open and encourage a free flow of communication. Also, the nuances of moderating content and keeping gaming platforms safe are immense, so they often rely on in-game moderation to address issues.

3. Their own gaming community isn't causing any red flags – yet.

4. Choosing a profanity filtering and content moderation platform is no easy feat. The market is flooded with options, many of which overpromise and underdeliver.

5. An actual full-blown reputation crisis like that some platforms have recently undergone isn't all that likely – they hope.

The Challenges

Gaming platforms are facing are shared by other forms of digital entertainment, and most platform developers want to protect both players and their bottom line. What's clear is that gaming companies, early-stage game developers, and forum and chat room moderators want a solution. That said, solving the problem is often easier said than done.

Strategies for Combating Hate Speech & Extremism

Community Guidelines

Early-stage game developers and gaming companies, along with forums, chat rooms, and moderators, are concerned about the severity of the problem. Instituting community guidelines is a good place to start.

In 2020, ADL and the Fair Play Alliance released the Disruption and Harms in Online Gaming Framework to help the industry better define and address hate and harassment in online games. The framework centers on four key elements of disruptive conduct in gaming:

- **Expression**

  What form does it take?

- **Delivery Channel**

  Where does it happen in and around online games?

- **Impact**

  Who is affected by it, and in what ways? What are the consequences?

- **Root Cause**

  Why does it happen? What does it express?

Gaming companies and platforms can adopt the framework, assess the efficacy of their current efforts to combat hate and harassment among users, expand their investments in staff and products, and make a regular practice of performing third-party audits to measure impact.

Filtering & Moderation

Of course, it's one thing to say you have a zero-tolerance policy against hate speech and extremism and another to put it into practice. While community guidelines help to outline expected codes of conduct, the ability to enforce those codes is where many platforms falter.

Ideally, content moderation should involve screening for, flagging, and removing inappropriate text, images, and videos that users post on a platform by applying pre-set rules for content moderation. Moderated content can include profanity, violence, extremist views, nudity, hate speech, hate raids, spam, and other forms of inappropriate or offensive content.

Some technology solutions have now evolved to a point where they can efficiently weed out the majority of unseemly content. Rather than solely relying on users themselves to report the misconduct, many gaming companies use a combination of proactive monitoring of public spaces, machine learning and reactive tools, and user reporting to monitor policy violations, moderate offensive behavior, remove troubling content, and suspend or ban offenders. Some platforms are also taking steps to address extremism, such as using artificial intelligence to detect offensive content.

The aim is to create something automated that works more like a referee: a system that can call out harassment right then and there when it happens, in real time.

Ways to Address Toxicity in Gaming

- Proactively detect objectionable content, flag it and identify the user sources, even before taking disciplinary action
- Introduce systems to reinforce good behavior and set community standards for how players should interact with each other in-game.
- Increase the level of transparency between developers and players.
- Overhaul reporting systems to more accurately punish behaviors that the player base considers to be toxic.
- Encourage parents to communicate with kids early and often to help them identify hate speech when they see it and know what next steps to take.
- Choose a profanity filtering and content moderation platform with childproofing game settings available.
- Punish behaviors considered to be toxic.
- Reward behaviors considered to be positive.
- Capture metrics and track trends for toxic content on topics and user communities.

Content Moderation & Filtering Critical Requirements

The need to reduce or eliminate inappropriate or unwanted content is becoming all the more urgent. The reality, however, is that not all solutions are created equal. For game developers who are committed to prioritizing advanced, automated profanity filters and content moderation capabilities,

The following 10 gaming platform capabilities are critical:

Expression

Perform content moderation (text moderation), image moderation, and video content moderation.

Scalability

Filter tens of thousands of messages per second on a single server.

Accuracy

Keep false positives and misses to a minimum and prevent users from bypassing filters with fast and accurate filtering in multiple languages to meet diverse business needs.

Pre-Approval/Rejection of Content

Allow moderators to pre-screen content in a pre-approval queue before it's visible to other users.

Multiple Application Use

Manage users and content across multiple applications, set up different filtering and moderation rules for each application and content source, and isolate moderators so they moderate content and users only for specific applications.

Ease of Use & Performance

Rapidly implement an integrated, developer-friendly solution that begins working quickly to ensure

Built-In Reporting Tools

Use content moderation and content filtering reports and analytics to better understand the user community, extract meaningful insights from user content, and improve overall business performance.

User Discipline

Automatically take action after a user's reputation/ trust score reaches a designated threshold, as well as use progressive disciplinary action to manage repeat offenders and enable moderators to escalate issues to managers immediately.

Security

Choose an on-premise solution that ensures each customer's data is isolated and ensures personally identifiable information (PII) is secure.

PII, COPPA, and EU Data

Provide strict compliance with COPPA, InfoSec, and EU Data Privacy Directive requirements.

Gaming Platform Content Moderation

The following five content moderation capabilities can position gaming platforms to protect their user community, save time and money, and vastly reduce the workload of human moderators:

Text, Image, & Video Filtering

Pre- and post-moderate text, images, and video in all gaming platform communications, including chat, forums, reviews, profile pictures, and other visible user account information.

User Flagging

Enable users to report content and images they deem inappropriate, resulting in content moderation and possible content removal or user suspension.

Blacklist Filtering

Use aggressive rules, natural language processing, and advanced algorithms to identify blacklisted words and phrases in both usernames and text.

Kids Chat Filtering

Reject any words or phrases that aren't included in the whitelist.

PII Filtering

Implement personally identifiable information (PII) filters, including email and phone number filters.

The Solution:

Intelligent Filtering & Advanced Content Moderation

Gaming is no longer just a hobby, and to describe it as a niche market is an understatement. It's a booming industry with wide room for growth.

As the gaming industry continues to experience rapid growth, platforms are coming to terms with the need to do everything possible to eliminate or at least significantly reduce hate speech and extremism. Most gamers want there to be consequences for bad online behavior, and it's becoming increasingly clear that in-game moderation isn't enough.

Gaming industry leaders around the globe need a solution that will identify and remove inappropriate or unwanted online text, image, and video content before it's visible to customers. They also need a solution that understands how extremism functions in online game spaces and that keeps pace with the myriad ways people find to skirt filters and barriers in order to express hate speech and extremist views. The most effective way to keep tabs on all of this content is by using profanity filters and advanced content moderation technology as a first line of defense.

cleanspeak

CleanSpeak is an enterprise-scale profanity filtering and content moderation platform that protects online communities from offensive and inappropriate language and images by preventing profanity and hate speech.

CleanSpeak filters billions of messages each month in real time. It's a flexible, cost-effective solution trusted by companies ranging from startups to Fortune 500 corporations, spanning a wide range of industries including gaming, entertainment, financial services, healthcare, education, and consumer goods. The intelligent filtering and advanced content moderation technology enables organizations to improve customer goodwill, reduce the risk of PR disasters, and preemptively save millions of dollars associated with negative publicity, lawsuits, and lost business.

For more than a decade, CleakSpeak has been advancing its profanity filtering technology to keep customer communications clean and productive and maintain a safe environment for users around the globe. For more than a decade, CleakSpeak has been advancing its profanity filtering technology to keep customer communications clean and productive and maintain a safe environment for users around the globe.

To see if CleanSpeak is right for your business, try it for free today.

cleanspeak

Combatting Hate & Extremism on Your Gaming Platform

cleanspeak.com