WHAT IS PROACTIVE VOICE MODERATION?

Unlike traditional moderation tactics such as reactive player reports, proactive voice moderation gives real-time awareness of the worst harms players are experiencing and empowers moderators to address toxic behavior holistically.

MODULATE

CONTACT@MODULATE.AI MODULATE.AI

Overview

This brief white paper from Modulate will cover:

1. How toxicity is harming game communities.
2. The prevalence of toxicity in games.
3. How studios can uncover toxic behavior.
4. Why proactive voice moderation is the future.

Toxicity in Games

Let's start by stating the obvious: modern online games and "metaverse" platforms have become far more social than their predecessors. While historically games focused on providing a specific curated experience to the player, they are now more akin to a "space" for players and their friends to congregate, from which the social group might choose a variety of experiences to partake in.

This shift, bolstered by the quarantine periods of the COVID-19 pandemic, has been a boon to many players. More than 75% of players report that video games help them to stay in touch with friends and connect with new people, and marginalized and at-risk demographics have reported that their online communities are a crucial safety net. Unfortunately, this shift towards social features also means that disruptive or harmful behavior can now more severely impact the experiences of a broader set of players than ever before. As such, online toxicity, including hate and harassment, has rapidly shifted from an already serious problem to a full-on crisis as games fight to keep their communities safe and inclusive for all.

A majority of multiplayer gamers (67%) say they would likely stop playing a multiplayer game if another player were exhibiting toxic behavior.

The State of Disruptive Behavior in Multiplayer Games (Unity)

Aren't player reports enough?

Reactive analysis of player reports is certainly important to give players a direct tool for improving the community, but it is unfortunately woefully insufficient to protect everyone. Less than ten percent of any kind of offense gets reported by players today, and for some of the most crucial offenses - such as child predators, violent radicalization, or influence towards self-harm - the victim almost never reports the offense.

Reactive moderation should be a piece of the puzzle, but not the whole strategy.

Prevalence of Toxicity

5 out of 6 adults ages 18-45 experienced harrassment in online multiplayer games.

Hate Is No Game: Hate and Harassment in Online Games (ADL)

It's important to understand that toxicity in gaming is everyone's problem. 83% of adult gamers report facing toxicity online, across every demographic of player, though often with emphasis on targeting the underprivileged. This prevalence has significantly impacted the public perception of gaming, as **80% of players believe the average gamer makes prejudiced comments while playing online.**

In reality, the vast majority of toxicity comes from a small contingent of repeat offenders, typically making up no more than 5% of the users on even the most unregulated platforms. But this small number of bad actors has a disproportionate impact.

And of course, not all toxicity is intentional - sometimes a generally supportive player has a bad day, misjudges the social norms of their community, or responds reflexively to a perceived threat, resulting in additional harmful behavior that must be handled separately. These players tend to be much more receptive than repeat offenders to education and support resources helping them learn to treat their communities more respectfully, underscoring how important it is to understand the nuance of toxicity on each platform to be able to respond to each instance appropriately.

Regardless of intent, though, negative behavior of any kind can instantly ruin a player's experience and impression of the game. Many players only need one negative experience to decide a game is not for them. This presents a serious dilemma to platforms. To protect their players properly, platforms must be able to reliably detect all forms of toxicity promptly, and take action swiftly enough to stop the harmful behavior while it's still happening, before the targeted player makes the call to simply leave entirely. But top games often have millions of conversations occurring simultaneously - how can they possibly identify and act on each instance of harm quickly enough?

Uncovering Toxicity

Discovering harm comprehensively and quickly is obviously a complex problem. It requires answers to qualitative questions like “what exactly counts as harm” and “how should we react to different types of harmful behaviors” that are widely debated among experts today.

But before we can even ask those questions, it first requires a method to identify, from amid all the conversations happening within the platform, which ones include toxic behavior. And for game teams to be able to respond quickly, these conversations must be flagged in real-time, while the problematic behavior is still ongoing. One way to identify these harmful conversations is to rely on other players filing reports against bad actors. Unfortunately, **moderation via reactive player reports does little to prevent toxicity**. Few players actually submit reports (between 5-25% of players depending on the title and genre), meaning that at least three out of four instances of toxicity will be missed. And since even one instance of serious toxicity is often enough to drive a player away, removing such a small fraction of bad actors may not actually decrease a game’s churn rate at all.

Further, some of the worst types of issues, such as child grooming or radicalization, will basically never be reported, as the victims in these cases are not in a state of mind to be able to recognize what’s happening or report it. And finally, player reports often are only submitted after the fact - meaning that even if they do result in punishments to bad players, the damage has already been done and the impacted players have probably already left the platform.

Another possible approach to randomly sample conversations from across a game or platform and dig into them further. This approach has some benefits – it’s possible to detect those more insidious harms using this method – but it’s certainly not reliable. It might find a decent chunk of the toxicity happening in the game, depending on how much data is sampled, but odds are that it will miss at least as much as player reports did (and can still be a major resource drain), albeit different types of things. It may appear obvious that an ideal selection process would be to simply dynamically shift focus to each relevant conversation right as toxicity begins to emerge. There’s only one proven way to do just that: proactive moderation. And while many proactive text moderation solutions are available, **only one proactive voice moderation solution exists today: ToxMod from Modulate**.

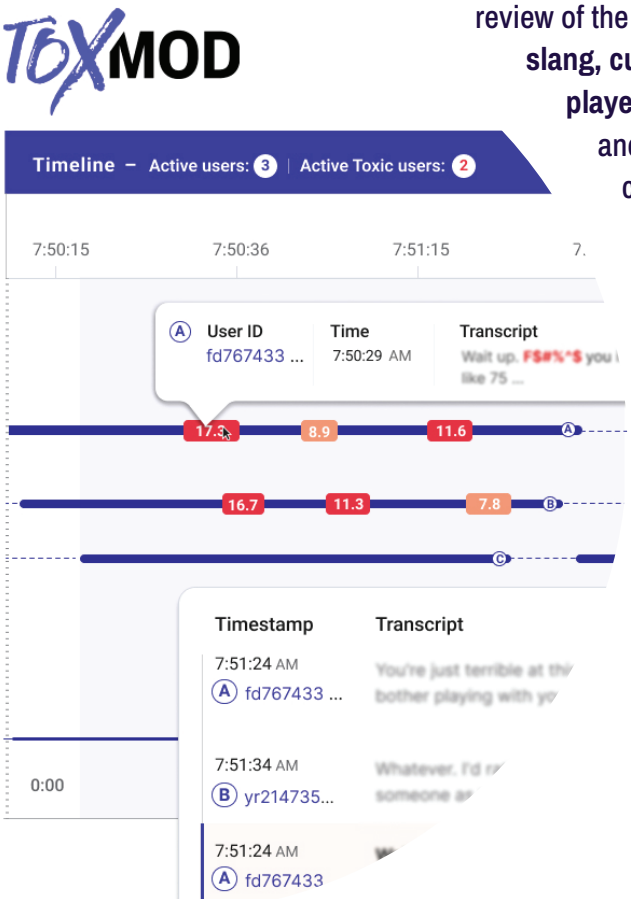
Preventing Toxicity
with Proactive Voice Moderation

It may sound like magic, but **ToxMod's secret sauce is its ability to focus dynamically on conversations as toxicity begins to emerge, without any manual direction.**

Instead of achieving this by analyzing every conversation in full detail (which would be prohibitively costly as well as an unnecessary privacy risk for players), ToxMod's patent-pending **proactive triaging models** use machine learning tuned continuously by moderators to quickly identify the key signals that show a conversation is taking a bad turn.

These triage models can't understand everything about a conversation at first glance, but they can look out for telltale signs of anger, distress, aggression, and even more subtle sinister intentions. What this means is that **Trust & Safety teams can actually focus on the specific conversations that are most important to review** amongst the millions happening at any point across the platform – without waiting for player reports or hoping that the activity will be caught with a random spot check.

From there, ToxMod's next-generation analysis engine can perform a deeper review of the situation, **bringing in additional understanding of context, slang, cultural norms, and the history between participating players** to ultimately provide complete context about the nature and severity of the offense and how it fits into the surrounding conversation.



Players are tired of the burden of community management being put on their shoulders, with their only options being often ineffective reports, muting and enduring, or leaving entirely. By being **proactive**, and showing players that **toxicity will not be tolerated**, the teams behind today's biggest games can show their players they take seriously their duty towards safe, inclusive, and positive online experiences, and **foster the healthy and engaging communities that players deserve**.

GET IN TOUCH TODAY