



Case Study

# Hearing the Problem: Modulate's Journey Bringing Real-time Safety to *Call of Duty*

Mariana Olaizola Rosenblat and Dorothee Baumann-Pauly

October 2025

# Contents

Learning Objectives .....	3
Case Synopsis .....	3
1. Introduction .....	4
2. Games as Social Platforms.....	4
3. Breeding Grounds for Extremism.....	5
4. From Christmas Code to Commercial Pilot: The Founder Narrative .....	6
5. Breakthrough with Activision.....	7
6. How ToxMod Works .....	7
7. Future Outlook .....	9
8. Questions.....	10
Endnotes.....	11

## How to Use This Case Study

This case is built for flexible use in the classroom, in workshops, or in professional training. Readers should review the story and learning objectives before the session, then use the discussion questions to explore the business, ethical, and regulatory dilemmas it raises. Instructors can adapt the material for group exercises, role-plays, or open conversation, depending on time and setting. The goal is to help participants think critically about how technology companies can balance safety, privacy, and business pressures.

## Learning Objectives

1. Diagnose the human rights risks associated with unmoderated voice chat —harassment, hate speech, grooming, and extremist recruitment—and articulate their real-world consequences.
2. Evaluate trade-offs between safety, privacy, and freedom of expression when designing or regulating real-time communications technologies.
3. Analyze the business case for and against adopting voice-chat moderation technology in games.
4. Propose metrics for determining whether voice moderation is having its intended impact on reducing toxicity and extremism in gaming platforms.
5. Formulate a business pitch from Modulate, or similar voice moderation providers, to online gaming companies.

## Case Synopsis

Exposure to hostile rhetoric and toxic harassment is commonplace in online gaming sites and, in some cases, leads to radicalization and real-world hostility. Despite mounting evidence that live voice chat is one of the primary vectors for the dissemination of extremist ideology<sup>1</sup> in games, most publishers have been reluctant to address the problem proactively, citing both player-privacy expectations and the technical complexity of moderating real-time communications.

In 2023, Activision partnered with Modulate, a Boston-based start-up, to deploy ToxMod, the first scalable, real-time voice-moderation system purpose-built for games. Early [results](#) published by Activision and Modulate in late 2023 were striking. From late August through October, the number of players exposed to severe instances of toxic voice chat dropped by about 50%, while retention among players in voice-moderated lobbies rose 28%.<sup>2</sup>

This case traces how Modulate's founders developed a mission-driven safety platform, the commercial and ethical hurdles they overcame to win Activision's business, and what their experience reveals about the wider challenges of governing user communications in immersive social platforms.

# 1. Introduction

Fictional Scenario: Four players are at a multiplayer voice chat lobby in *Call of Duty*:

- ShadowEagle (an adult recruiter for the white supremacist group Patriot Front)
- BlasterKid12 (age 13)
- NovaStorm (a young Hispanic player)
- L33tSniper (a silent observer)

ShadowEagle compliments BlasterKid12's gameplay and builds rapport by validating BlasterKid12's frustration with school and adults. The conversation shifts to vague political and societal commentary as ShadowEagle suggests that "things are broken" and that schools and media don't tell the full truth. He introduces coded conspiracy theories about immigrant invasion and White racial erasure.

When NovaStorm challenges the conspiracy theories, ShadowEagle detects his Hispanic accent, shouts a racial slur and tells NovaStorm to "shut up and go back to where you came from." NovaStorm abruptly leaves.

ShadowEagle tells BlasterKid12 they seem mature for their age and invites them to a private Discord server for "real discussions." He reads the invite link aloud to avoid text-based moderation and urges secrecy, suggesting that adults wouldn't understand. BlasterKid12 expresses interest and notes the server name.

## 2. Games as Social Platforms

With a global annual turnover of nearly \$200 billion, the video game industry generates more revenue than the movie and music industries combined.<sup>3</sup> Over three billion people worldwide play online games, and a majority of them play multiplayer games, in which players located in different parts of the world can interact and communicate with one another in real time.<sup>4</sup>

Online multiplayer games appeal to large audiences because, in addition to providing entertainment, they offer venues for networking and community building. Most multiplayer games today allow participants to communicate live via in-platform text and voice chat functionalities. Such communication tools have become staple features of many game titles and key drivers of userbase growth and retention. Some players also turn to Discord, a popular gaming-adjacent platform that offers expanded chat and long-term networking functionalities.

The social nature and popularity of online gaming spaces helps explain why they have become attractive venues for extremists, including those looking to recruit and mobilize for violence. Gaming sites offer easy access to large numbers of youth, many of whom are unsupervised, highly impressionable, and yearning for connection. Brief encounters in individual matches in Activision's *Call of Duty*, or longer contact via the 3D-immersive spaces of *Minecraft* and *Roblox*, can lead to invitations to join more private online venues

in applications like Discord and Steam, the largest storefront for digital games, which also offers community forums. It is often in these gaming-adjacent spaces, as well as on fringe sites frequented by extremists, that longer-term indoctrination and mobilization occur.

### 3. Breeding Grounds for Extremism

As modern video games are designed to encourage and facilitate social interaction, the unfortunate challenge is that both good and bad actors have equal access to these tools for building friendships. In particular, while the scale or prevalence of hostile rhetoric and behavior in online games remains understudied, there is enough evidence from surveys and anecdotal reports to ascertain that extremists use in-game chat functions to express, enact, and normalize their ideologies. According to a 2023 representative [survey](#) commissioned by the NYU Stern Center for Business and Human Rights, 51% of gamers across five major markets—the United States, United Kingdom, Germany, France, and South Korea—encountered statements inciting discrimination or violence while playing multiplayer games in the previous year. These included statements supporting the idea that a particular race or ethnicity should be eliminated and that violence against women is justified.<sup>5</sup>

#### Encounters with Extremism

Percentage of respondents who came across statements supporting the use of physical violence against a particular person or group based on their identity:

United States of America	35%
United Kingdom	25%
South Korea	30%
France	25%
Germany	29%

Percentage of respondents who came across statements portraying a particular ethnic, gender, or religious group as inferior:

United States of America	41%
United Kingdom	31%
South Korea	50%
France	23%
Germany	33%

Percentage of respondents under 18 who came across statements expressing support for the idea that:

The white race is superior to other races	16%
A particular race or ethnicity should be expelled or eliminated	17%
Using violence is justified or necessary to achieve a political aim	15%
Women are inferior	18%
Violence against women is justified	6%

Percentage of respondents 18 and over who came across statements expressing support for the idea that:

The white race is superior to other races	13%
A particular race or ethnicity should be expelled or eliminated	16%
Using violence is justified or necessary to achieve a political aim	12%
Women are inferior	21%
Violence against women is justified	10%

Extremism via in-game chats also takes the form of severe harassment. In NYU's multi-national survey, 36% of participants had experienced some form of extreme or hate-based harassment (e.g., stalking, hate-raiding,<sup>6</sup> sexual harassment, violent threats, doxxing,<sup>7</sup> or swatting<sup>8</sup>) while playing online multiplayer games in the last year. Further, the Anti-Defamation League has found that a considerable portion of this harassment is identity-based, targeting women, African Americans, Asian Americans, Jews, and LGBTQ+ players.<sup>9</sup>

Compared to text chat, in-game play, or user-uploaded content such as images and videos, voice chat is a main vector for toxic behavior, with 49.4% of surveyed players in one study saying that they have experienced toxic incidents via voice chat.<sup>10</sup> Nevertheless, more than two-thirds of online game players use voice chat, according to industry research.<sup>11</sup> Further, in many cases, voice chat is effectively non-optional, such as when a player needs to coordinate with teammates in a fast-paced competition. As such, curbing toxicity and extremist behavior in voice chat has emerged as a key challenge for the industry.

## 4. From Christmas Code to Commercial Pilot: The Founder Narrative

In early 2018, Mike Pappas and his MIT-educated co-founder left their day jobs and spent six frenetic months cold-calling studios, from tiny indies to giants like Riot and Activision, pitching a product that would allow players to customize the sound of their voice in a game. The response surprised them. Rather than flashy voice cosmetics, or accessories that modify the appearance of a game character without affecting their game performance, studio executives wanted help solving “toxicity.”

**“My co-founder wrote the first line of code for what became Modulate on Christmas Day 2015.”**

— Mike Pappas, CEO, Modulate

The studios' first request was: *Could Modulate make every female player sound male?* This way, women, who disproportionately suffer harassment in gaming platforms,<sup>12</sup> would be shielded while still being able to participate in real-time communications. But Pappas told this study's author that he recoiled: “There must be a better way than telling innocent players to hide who they are.” His team began repurposing their models to detect toxicity instead of disguising identity. Thus, ToxMod was born.

**“Game studios were saying, ‘Yes, I need this... People are unwilling to participate in voice chat because of harassment,’ and that hurts retention.”**

— Mike Pappas, CEO, Modulate

Studios hesitated at first. They worried that players would see ToxMod as an infringement of their privacy and no longer use in-game live chat or, worse, leave the game altogether. Another barrier loomed even larger: sticker-shock. Pappas explains: “You can process millions of tokens of text for the cost of a single hour of audio.” Because the benefits of large integrations (retention, brand value) materialize months after implementation, game company CFOs demanded proof before funding them, creating a classic “catch-22.”

Yet studios continued to express interest in toxicity detection, mostly out of concern for player retention. In addition, from 2019 on, a series of high-profile incidents making use of gaming platforms or terminology—including the racist-inspired mass shootings in Christchurch, New Zealand,<sup>13</sup> Buffalo, New York,<sup>14</sup> and El Paso, Texas<sup>15</sup>—heightened the external pressure on gaming companies to address toxicity and extremism on their platforms.

## 5. Breakthrough with Activision

In late 2022, a product manager from Activision connected with Modulate saying, *It seems like you're doing something about toxicity. Let's chat some more.* After weeks of technical deep-dives—and no airtight ROI model—Activision paid for a live experiment inside the company's flagship first-person shooter game, *Call of Duty*.

Within weeks, the data was unequivocal: 25% to 33% less toxicity exposure during the two-month trial period and a 28% increase in player retention on voice-moderated venues compared to non-moderated ones.

**“They put real money on the line for an experiment.”**

— Mike Pappas, CEO, Modulate

Activision green-lit a wider<sup>16</sup> rollout (excluding Asia Pacific) and, crucially, agreed to let Modulate publish the results, elevating the unknown start-up to a company that powers *Call of Duty*. The company proudly noted in its anti-toxicity progress report for 2023: “*Call of Duty* is taking the next leap forward in its commitment to combat toxic and disruptive behavior with in-game voice chat moderation.... *Call of Duty*'s new voice chat moderation system utilizes ToxMod, the AI-Powered voice chat moderation technology from Modulate, to identify in real-time and enforce against toxic speech—including hate speech, discriminatory language, harassment and more.”<sup>17</sup>

After ToxMod's international deployment, *Call of Duty* saw a roughly 50% reduction in players exposed to severe instances of disruptive voice chat since *Modern Warfare III*'s launch.<sup>18</sup> With the backing of a high-profile game studio, Modulate gained global recognition. Speaking invitations, press coverage, and a wave of game studio prospects followed.

## 6. How ToxMod Works

ToxMod is a voice-native machine learning system, meaning that the AI-powered detection models are trained on 18 different languages from around the globe to analyze voice clips directly, rather than transcribing and analyzing text. Non-voice native systems, which require such transcription, take time and large amounts of computing resources, and lose the layers of nuance and audio context which are critical to understanding the meaning behind words. This is especially true in the gaming context, where trash talk is endemic and it is important to differentiate between competitive, enjoyable banter and hostile behavior. Upon detecting potentially problematic interactions, with the specific threshold determined by each game's Code of Conduct or “community guidelines,” ToxMod quickly organizes these incidents according to the level of risk and escalates them in real time to company representatives for their swift examination and resolution.<sup>19</sup>

## ToxMod's Role in Live Voice Moderation

In the introductory fictional scenario, Modulate's ToxMod is active during the in-game lobby conversation. ToxMod continuously analyzes live voice chat for harmful behavior, including hate speech, harassment, grooming, and hateful language.

As ShadowEagle builds rapport with BlasterKid12 and introduces coded ideological references, ToxMod detects a progression of concerning signals:

- Early grooming behavior through targeted flattery and attempts to build trust with someone who appears to be a younger player.
- Escalating language indicating exposure to extremist ideas, including conspiracy-related codewords and attempts to shift the conversation off-platform.
- Behavioral patterns consistent with ideological grooming and radicalization tactics seen in other flagged incidents.

ToxMod flags the conversation in real time for review, assigning a high severity score based on the combination of:

- User age mismatch
- Attempted off-platform migration
- Escalating ideological content

A trust-and-safety moderator at Call of Duty is alerted with a recording and transcript of the flagged segment. Meanwhile, the system initiates a temporary communication restriction on ShadowEagle to prevent further harm while human review is underway.

By identifying intent and context—not just keywords—ToxMod helps surface subtle but dangerous interactions that often evade detection through manual reporting or text-based filters alone.

ToxMod is underpinned by a philosophy regarding the appropriate balance between the need to counter toxicity online and players' reasonable expectations of privacy. As explained by CEO Mike Pappas, the architecture of Modulate's moderation system relies on "machine learning triage," a strategy that involves having the AI listen to conversations from a distance, so to speak, without recording and transcribing every word that every player says. Only when the AI picks up early signals of hostility or vulnerability does it listen more closely and carry out a contextual analysis, which it then sends to the company.

These early signals can be any of a large number of possibilities, trained into Modulate's AI through hundreds of millions of hours of examples, but Pappas notes a few examples: whether someone is shouting or crying, whether some people in the conversation have

become uncharacteristically quiet, or whether a younger child is being pulled into a one-on-one conversation with an unconnected and suspicious adult. This approach reflects Pappas' playground analogy for digital games, in which parents or supervisors let their children explore while staying attentive enough to intervene if there are warning signs of danger, thereby striking a healthy balance between privacy and safety.<sup>20</sup>

Added to this architecture is Modulate's commitment to anonymizing all user data and to never selling or renting that data.<sup>21</sup> On Activision's side, the publisher has also made adjustments to its terms of service and user agreement to inform players that their voice communications may be monitored in accordance with the new policies.<sup>22</sup>

## 7. Future Outlook

Now that Modulate has the data to prove its product's success, additional game studios are considering ToxMod's deployment. But cost continues to act as a barrier. Pappas explains: "No one is saying, *We don't care about our players. Let's just let them suffer.* But they're asking themselves, *Is there a cheaper way to solve this problem?*" Yet, according to Pappas, studios have not had much success in doing so "because this is such a complicated problem that really needs dedicated experts."

Regarding privacy, concerns over eavesdropping persist, but the industry consensus has shifted from *if* to *how* voice can be monitored responsibly. "It's always a concern, but it's not a blocker," says Pappas. "Everyone understands you can do this in a privacy-safe way; they just want to make sure we're thoughtful."

Alongside continuing to pitch their technology to additional game companies, Modulate is partnering with popular tools that game developers already use to power voice chat, such as Vivox, Agora, Photon Voice, and even Discord. By building these partnerships, Modulate makes it possible for studios to quickly plug ToxMod into their existing systems and immediately see how it helps detect and reduce harmful behavior during voice chats in games.<sup>23</sup>

In parallel, regulatory momentum in the EU,<sup>24</sup> UK,<sup>25</sup> Australia<sup>26</sup> and across US states<sup>27</sup> has shifted boardroom calculus from *Is this necessary?* to *Can we afford not to?* Activision's case suggests that investments in safety can pay for themselves through higher engagement and reduced churn. But even companies that remain unconvinced by the business case may soon no longer have a choice under evolving regulations which increasingly require proactive measures to combat illegal content and serious harm.<sup>28</sup>

## 8. Questions

- 1** To what extent do gaming platforms have a responsibility to address hateful rhetoric and behavior that takes place on their platforms?
- 2** To what extent should game companies monitor players' text and voice conversations to detect potentially extremist rhetoric and behavior?
- 3** Should players have a reasonable expectation of privacy with regard to the content of their speech on gaming platforms?
- 4** How can game companies balance the need for healthy gaming communities and the privacy interests or preferences of their user base?
- 5** What are the benefits and drawbacks of deploying AI-powered moderation tools in gaming platforms?
- 6** How can third parties motivate game publishers and platforms to adopt proactive moderation technology?
- 7** How can third parties, such as academic researchers, motivate game publishers and platforms to allow them access to communications data for the purposes of research and policy development?
- 8** What can we learn about the business case for human rights from this case study?

## Endnotes

- 1 There are various understandings of the term “extremism.” In this case study, extremism refers to a belief system held together by unwavering hostility towards a specific “out-group.” For more details on the definition, see J. M. Berger. *Extremism*. United States: MIT Press, 2018.
- 2 Modulate and Activision, [The Impact of AI Voice Moderation on the Call of Duty Player Experience: A Case Study](#).
- 3 NewZoo, [The global games market will generate \\$187.7 billion in 2024](#).
- 4 Fabio Duarte, [How Many Gamers Are There? \(New 2025 Statistics\)](#), July 2025.
- 5 NYU Stern Center for Business and Human Rights, [Gaming The System: How Extremists Exploit Gaming Sites And What Can Be Done To Counter Them](#), May 2023.
- 6 Attacking someone in coordinated fashion to overwhelm them with insults and threats.
- 7 Publishing someone’s personally identifiable information such as their address for the purpose of intimidation.
- 8 Prank calling law enforcement to someone’s location; the term refers to the dispatching of police SWAT teams.
- 9 ADL, [Hate Is No Game](#), 2022.
- 10 As opposed to text chat, in-game play, and user-uploaded content such as images and videos. Speechly, [Gamers Say They Love Voice Chat, But 49% Have Been Victims of Toxic Behavior](#), March 2023.
- 11 Speechly, [Gamers Say They Love Voice Chat, But 49% Have Been Victims of Toxic Behavior](#), March 2023.
- 12 Women in Games, [Over Half Of Women Gamers Experience Online Abuse](#), April 2023
- 13 Graham Macklin, [The Christchurch Attacks: Livestream Terror in the Viral Video Age](#), July 2019.
- 14 Office of the New York State Attorney General Letitia James, [Investigative Report on the role of online platforms in the tragic mass shooting in Buffalo on May 14, 2022](#), October 2022.
- 15 Evan Urquhart, [The Real Connection Between Video Games and Mass Shootings](#), August 2019.
- 16 It was deployed in all regions except for Asia Pacific.
- 17 Activision, [Call Of Duty Takes Aim At Voice Chat Toxicity, Details Year-To-Date Moderation Progress](#), Update 11/10/2023.
- 18 Modulate and Activision, [The Impact of AI Voice Moderation on the Call of Duty Player Experience: A Case Study](#).
- 19 ToxMod, [Proactive harm detection for real-time conversations](#); Modulate, [What is proactive voice moderation?](#)
- 20 Mike Pappas, [Social Safety in Games: Moderating Voice Chat in the Metaverse](#), October 2023.
- 21 Modulate, [ToxMod for gaming](#).
- 22 Activision, [Terms of Use](#), November 2022.
- 23 See, e.g., [Modulate Expands Voice Chat Safety with Discord Social SDK Support](#).
- 24 [EU Digital Services Act](#), 2022.
- 25 [UK Online Safety Act](#), 2023.
- 26 [Australia Online Safety Act](#), 2021.
- 27 Center for Social Media and Politics, [The State of State Technology Policy: 2024 Report, December 2024](#).
- 28 See generally, NYU Stern Center for Business and Human Rights, [Online Safety Regulations Around The World: The State of Play and The Way Forward](#), April 2025.

NYU Stern Center for Business and Human Rights  
Leonard N. Stern School of Business  
44 West 4th Street, Suite 800  
New York, NY 10012  
+1 212-998-0261  
bhr@stern.nyu.edu  
bhr.stern.nyu.edu



Center for Business  
and Human Rights

GENEVA CENTER  
FOR BUSINESS  
& HUMAN  
RIGHTS