Keywords Studios
Imagine More

MODULATE

ActiveFence

Take this.

The Future of Content Moderation in Gaming:

A Unified Approach with AI and Human Touch

Contents.

Introduction.

Today, most people recognize the importance of content moderation in building safe, healthy, and thriving online gaming communities. Indeed, players themselves now recognize the negative impact of toxicity in gaming and consider effective content moderation not just desirable, but essential in the current digital landscape. A 2023 report on toxicity by Unity revealed that of over 2,500 survey respondents, only 9% were unwilling to be recorded for the purposes of detecting and preventing toxicity in their game. In the same report, 81% of multiplayer gamers agree that protecting players from toxic behavior should be a priority for game developers.

But why should we be paying attention to moderating content in video games to begin with? We respond with a question of our own: Are popular multiplayer games like Fortnite and Roblox simply "games," or are they more akin to social platforms? These games are played by millions and are increasingly competing with "conventional" social platforms for users' time and attention. Statista estimates 70.2 million Daily Active Users (DAU) in Roblox for Q3 2023, while Fortnite reported a record-breaking 100 million players in the month of November 2023. The player bases of either of these hugely popular games are about one-third of the total estimated 353 million active users on X. Clearly games are an important part of the virtual social ecosystem and players should be protected from harm and toxicity, just as we would expect people in physical spaces to have access to safe and enjoyable experiences.

There are three crucial reasons why video game developer studios must prioritize effective content moderation:

1. **Content moderation is crucial for the wellbeing of our society**
2. **New legislation makes content moderation non-negotiable**
3. **Moderation is good for business**

Content moderation is critical for the wellbeing of our global society

The positive effects of gaming are well documented – a 2021 study from the Oxford Internet Institute found that "Players who objectively played more in the past two weeks also reported to experience higher well-being." However, toxicity in gaming communities is also well known – the ADL's 2023 annual survey of toxicity in gaming showed that, "though still exorbitantly high levels of harassment exist for all game players, in 2023 harassment against adults declined overall and across every category; concerningly, harassment of teens and pre-teens increased across nearly every category."

75% Three-quarters of teens and pre-teens experienced harassment in online multiplayer games in 2023.

Toxicity can mean harmful behavior like racial or sexual harassment, bullying, cultural hate speech, and more, but other forms of toxicity can be even more nefarious.

The most dangerous user-generated content is the kind that transcends the digital world into tangible harm, like doxxing, child exploitation, content promoting self-harm or suicide, and threats of terrorism or violence.

Enacting content moderation is no longer a matter of choice but is instead a societal obligation that allows us to pivot away from the harmful and towards a safer, more inclusive online world.

New legislation makes content moderation non-negotiable

In recent years, countries around the world have enacted online safety laws and regulations, aimed at ensuring safety in the virtual and physical world.

Two examples of this trend are the Digital Services Act (DSA) and the UK Online Safety Act. These have imposed strict guidelines that make online gaming platforms, such as massively multiplayer online games (MMOs) and social role-play games, liable for the content shared in-game. Besides the demand to moderate illegal content like hate speech, child abuse, exploitation, and disinformation, game developers are now required to conduct regular risk assessments, publish clear policies, and produce annual transparency reports detailing their moderation actions. Later this year, UK regulators will be drafting new Codes of Practice, specifically around handling illegal content and providing protections for children, which game developers must also adhere to, in order to comply with these new laws.

Regulation spotlight

Interestingly, the UK Online Safety Act highlights the importance of combining proactive technology with human intervention to achieve a balance between protecting online users and maintaining their freedom of expression.

The regulations have a clear focus on identifying and removing illegal content. But perhaps just as importantly, they necessitate the creation of transparent, robust, and efficient content moderation processes to ensure that users have safe experiences where the enforcement of rules is consistent. Neglecting these regulations can result in substantial fines and reputational damage, making compliance a business priority.

Let's get real: Content moderation is good for business

Toxic players alienate your core demographic, driving away potential revenue from future customers. Researchers Rachel Kowert, Ph.D., and Elizabeth Kilmer, Ph.D., found that 6 out of 10 players reported that they decided not to spend money in a game because of how other players treated them in that community. Likewise, in Games as Social Platforms, Constance Steinkuehler from the University of California Irvine revealed a 54% revenue gain for games that remove toxic content from their platform.

By prioritizing creating welcoming and safe online communities, these studies suggest that you will in turn increase player engagement, player retention, and profitability.

54% revenue gain for games that remove toxic content from their platform.

The key to creating healthier digital spaces

It's clear that moderation is necessary for any platform that supports user-generated content. But determining the most effective approach for moderation remains a crucial question.

Some argue that artificial intelligence alone is the ultimate solution to moderate UGC at scale, while others believe that only human decision-making can suffice, especially when faced with difficult and nuanced moderation decisions.

In this paper, we propose a third, and superior, path — a combination of AI and human intelligence.

We'll explore the benefits and challenges of combining artificial intelligence and human intelligence (AI + HI) in your content moderation workflow. You'll discover how AI and HI cooperate to improve efficiency, accuracy, and scalability, resulting in optimal moderation outcomes for your platform and your players.

HUMAN INTELLIGENCE

The secret superheroes of the internet.

Content moderation is, at its heart, an undeniably human endeavor. And for most of the internet's history, this role has been predominantly carried out by human beings.

In the early days of the internet, community volunteers often took on the role of moderating and upholding the rules. But as the internet grew, big social platforms and online games realized the importance of a dedicated moderation team, which led to the creation of paid moderation roles and services that we are familiar with today.

Still, platforms continued to rely almost entirely on humans to keep their spaces safe and secure. In time, some organizations began to apply rudimentary technology, like allow/disallow lists (a limited list of words or phrases that are automatically approved or rejected based on a code of conduct) to assist moderators in their work. These would later be joined by advanced AI technologies that use machine learning to detect violative content in text, images, video and audio.

But while technologies evolved, most moderation workflows continued to depend on humans to review, analyze, and approve or reject content without intervention or assistance from technology.

Why humans excel at content moderation

There is a reason that humans are at the center of content moderation: They are very, very good at it.

Human comprehension thrives on context. A single word, image, or meme can have different meanings across various situations, cultures, and communities. For example, the phrase "I am going to murder you, you can't hide any longer" could be appropriate banter in a first-person shooter game, yet entirely out of place in a forum dedicated to the same game. Here, the expertise of human moderators becomes apparent. As humans, our skill lies in interpreting the many subtleties of human expression and intent.

We are also incredibly adaptable. Moderation isn't a static task – it's as dynamic as the digital landscape itself. Language changes, new trends emerge, and cultural and geopolitical shifts happen — consider the lightning-fast emergence of new hashtags and TikTok challenges. Humans (sometimes with the help of a quick Google search) can keep up with this mutability, adjusting moderation tactics and strategies as needed.

Additionally, humans are good at assessing complex scenarios and connecting the dots using intuition, empathy, and innate human knowledge. For example, there may be situations where content doesn't explicitly violate any guidelines but is still inappropriate due to the underlying implications or the specific combination of elements. Some of the most dangerous content — like child grooming and insidious bullying — can fall under this category.

Human moderators have been integral to content moderation ever since the inception of the internet, and for good reasons: When it comes to human interactions, they offer indispensable understanding, flexibility, and empathy.

Why moderators are the secret superheroes of the internet

Like superheroes, moderators show up every day and help people in need — typically without any recognition of their invaluable (yet often silent) services. Their role in making the internet a safer place couldn't be more crucial. Here are three reasons that moderators are the secret superheroes of the internet:

- **They have extraordinary skills**. Moderators are equipped with sharp analytical skills, deep understanding of community guidelines, and they excel at making difficult decisions quickly, allowing them to swiftly assess and manage potential threats or harmful content flagged by technology.
- **They save lives.** Moderators carefully examine content that may be dangerous or pose serious threats. They act swiftly to mitigate harm, safeguarding players from injury — and sometimes even providing real-life protection.
- **They are fearless.** Moderators courageously face internet challenges head-on. They're essential in dealing with disturbing content and handling sensitive situations, showcasing their bravery and resilience.

The emotional toll of content moderation

There is an unfortunate paradox at the heart of content moderation. While humans offer a unique and specialized skill set required for the job, they are not shielded from the psychological harm brought on by sustained exposure to explicit, toxic content.

The role of a content moderator can be profoundly challenging and demanding. These professionals confront explicit, graphic, and disturbing content on a daily basis that can significantly erode their mental health and overall wellbeing— an ugly reality that has, tragically, been ignored for years.

In Content Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional Study, researchers from Middlesex University in London found that content moderators "have high rates of psychological distress and secondary trauma, and lowered wellbeing."

In this study and others, researchers have drawn parallels with first responders such as police officers and EMTs who regularly encounter indirect trauma. Content moderators also find themselves wrestling with an enhanced risk of secondary traumatic stress (STS) and vicarious trauma symptoms. Chronic exposure can lead to mental health issues, spanning anxiety, depression, significant emotional burnout, and in severe instances, even post-traumatic stress disorder (PTSD).

As noted in researcher Sarah T. Roberts' landmark study Behind the Screen: Content Moderation in the Shadows of Social Media:

> **"'[Content moderation is] permanently damaging.' The effects of that damage can be even more powerful when workers report an inability to sufficiently separate the responsibilities of their jobs from their time off the clock, whether it was sufficiently divorcing their sense of protecting users from seeing or experiencing harm, or the phenomenon of something disturbing from their workday invading their psyche when at home."**

The role also impacts how content moderators see themselves. In an interview with Harvard Business Review, Roberts notes that moderators will often refer to themselves as "a janitor or a trash collector — people who deal with the refuse, the detritus."

Additionally, job challenges include using outdated systems (which may force moderators to see harmful content for longer than necessary or in more emotionally damaging formats/contextualization due to tech lag), aggressive productivity and accuracy metrics, and poor working conditions coupled with low wages.

Similar sentiments have been echoed elsewhere. In a series of round tables hosted last year by Take This, they found burnout among community managers to be a near universal experience. As discussed by one participant,

"...I think working in this space burns people out to be honest… I spent the past week in workshops reading nothing but horrific slurs… and needed to take PTO after that, and rightfully so." Several also noted that, due to the psychological strain of the job and consequent burnout, the lifecycle of a community manager is three to five years. As discussed by one participant, "Community managers… are the sin eaters of the industry. There is no support there, emotional or otherwise. So if you're supposed to be this toxicity sponge, add to the fact that in our industry people only last three to five years before they are physically or mentally unable to do it anymore."

Knowing this, we have an unequivocal ethical obligation to extend the same level of empathy and care toward human moderators that we typically reserve for other first responders.

In The Psychological Impacts of Content Moderation on Content Moderators: A Qualitative Study, another study from Middlesex University, researchers suggest that "... it is crucial for companies to provide psychoeducation, intervention and trauma-informed care [for content moderators]".

In Content Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional Study, researchers found that "supportive work environments where moderators are given opportunities to create strong collegiate networks and are provided feedback about how important their work is would help ameliorate the relationship between their exposure to distressing content and the potential adverse effects."

"Additionally," they write, "there is some evidence that a work culture which encourages healthy working practices through taking breaks, would also help the mental health of content moderators. These results underscore the importance of organizational support and the implementation of policies and practices that prioritize [content moderators'] well-being."

Mounting legal challenges for gaming and social platforms

Content moderators have successfully sued several prominent online platforms, claiming that the companies failed to protect their wellbeing by not properly mitigating their exposure to harmful content.

In a recent ruling, a Barcelona-based company was found to be responsible for psychological damage suffered by a content moderator.

The findings in these studies underscore how important it is for organizations to use every resource at their disposal to champion the mental wellbeing of their content moderators. Unfortunately, platforms have historically failed to support moderators in this way. In her HBR interview Content Moderation is Terrible by Design, Sarah T. Roberts discusses the difficulty of implementing what appears to be a basic wellbeing measure, like providing extra breaks without any penalties. “A lot of companies pay lip service to [that practice],” she said. “But many moderators have told me they don’t seek support because it means letting their boss know they’re having difficulty with a central function of their work.”

Good news about moderator wellbeing

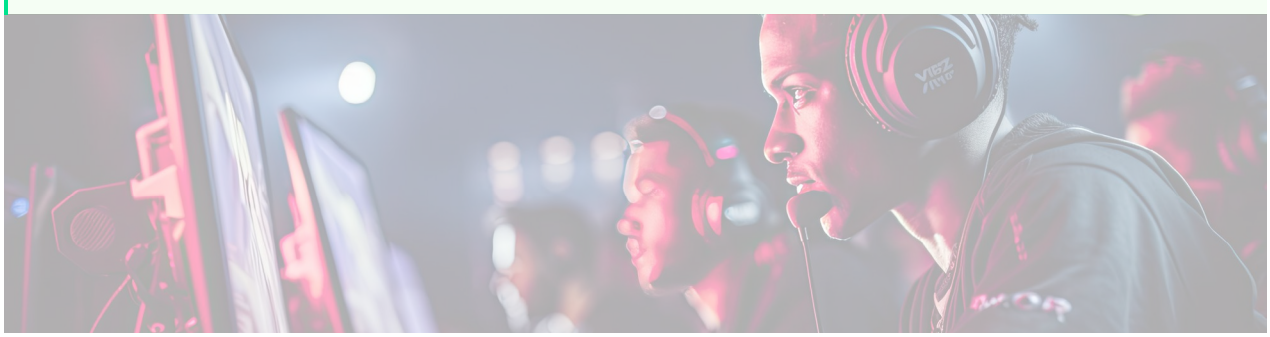
As research institutions, tech solution firms, and providers of human moderation services increasingly advocate for the protection of moderators from harm, the industry is taking notice.

IN 2023, the Atlantic Council’s Democracy + Tech Initiative at the Digital Forensic Research Lab (DFRLab) launched a multi-sector Task Force for a Trustworthy Future Web, uniting 40 experts spanning industry, civil society, and other sectors.

In its comprehensive report Scaling Trust on the Web, the task force shares its 8 key recommendations for better online spaces. Crucially, key finding #3 is this:

Protecting healthy online spaces requires protecting the individuals who defend them.

Implementing workplace wellness programs to address the needs of those exposed to harmful content is paramount; aside from the stated health impacts on moderators, platforms may face liability and a decrease in productivity if they do not make long-term investments to protect their employees… Interventions such as image blurring and moderation tools can help improve experiences for human moderators.



Platforms need to actively offer programs that enhance wellbeing and resilience, ensuring moderators feel a deep sense of autonomy, collaboration, and purpose.

As it turns out, AI technology is another effective and often underutilized instrument for taking preemptive protective measures.

Many platforms have adopted AI-driven content moderation solutions to help prevent players from encountering distressing or damaging content. But these solutions have another purpose that is equally important — they play a crucial role in shielding content moderators from excessive and unnecessary exposure to explicit material.

As we'll examine later in this paper, with AI, platforms can automatically and efficiently detect, analyze, and remove the most explicit content (think pornography, unambiguous hate speech, or known Child Sexual Abuse Material [CSAM]) before it reaches moderators, significantly reducing the volume of disturbing content that digital responders need to review.

In fact, ActiveFence research reveals that 80-90% of harmful content can be automatically detected and addressed without moderators needing to review it. Moreover, the majority of the remaining content is frequently generated by a small group of users, indicating that user-level moderation could effectively reduce the number of items viewed - minimizing the emotional toll on moderators and improving their efficiency.

(It should be noted here that the nuanced nature of human communication and the evolving landscape of online interaction means that, regardless of these stats, AI will never replace the need for human moderators. Beyond handling complex content, humans will always be needed to develop moderation strategies and advocate for essential safety measures.)

AI technology not only protects moderators from direct exposure to harmful content, but it has another vital purpose: With AI, moderators can save lives by identifying serious threats that require immediate attention.

Why the future of moderation requires more than human intervention

With the gaming audience constantly expanding, platforms today face a new reality: Human intelligence alone is no longer enough to enforce community norms.

"We've entered a new age," says Sharon Fisher, Global Head of Trust & Safety at Keywords Studios. "It's imperative that technology and humans work together — not just to keep online communities healthy, but to protect our most vulnerable players from real-life threats."

168%

An alarming rise in real-life threats

Between 2021 and 2023, moderators at Keywords Studios have recorded a 168% increase in instances of CSAM, violent threats, and self-harm/suicide-related content.

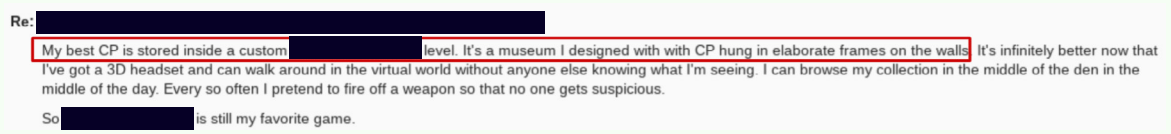
The surge in online threats like CSAM, terrorism, violence, and self-harm or suicide content has amplified the need for timely identification and escalation to authorities.

Unlike human moderators who have limited capacity, AI can detect and flag this dangerous content in real-time, drawing attention to time-sensitive issues swiftly and efficiently.

Below are just two disturbing examples of child predators using video games to share CSAM and communicate with children.

Example: Predator abuse of in-game features

Child predators are notoriously innovative, constantly seeking ways to misuse platform features to perpetuate their abuse. In one example, ActiveFence found child predators using in-game features to create “virtual museums” where they store and view their CSAM collection. These individuals used virtual reality headsets for a more realistic experience.

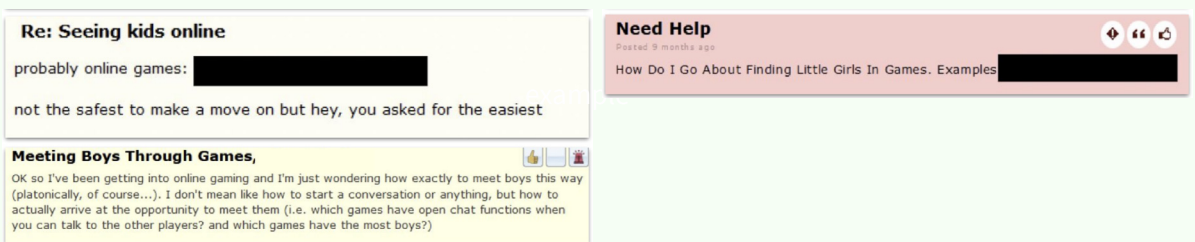

Re: [redacted]

My best CP is stored inside a custom [redacted] level. It's a museum I designed with with CP hung in elaborate frames on the walls. It's infinitely better now that I've got a 3D headset and can walk around in the virtual world without anyone else knowing what I'm seeing. I can browse my collection in the middle of the den in the middle of the day. Every so often I pretend to fire off a weapon so that no one gets suspicious.

So [redacted] is still my favorite game.

Example: Grooming in gaming platforms

Multiplayer games offer child predators a wide range of opportunities to interact with minors. By monitoring predator communities in the dark web, ActiveFence was able to intercept several conversations between child predators, discussing ways to use such games to interact with children.


Re: Seeing kids online

probably online games: [redacted]

not the safest to make a move on but hey, you asked for the easiest

Meeting Boys Through Games,

OK so I've been getting into online gaming and I'm just wondering how exactly to meet boys this way (platonically, of course...). I don't mean like how to start a conversation or anything, but how to actually arrive at the opportunity to meet them (i.e. which games have open chat functions when you can talk to the other players? and which games have the most boys?)

Need Help

Posted 9 months ago

How Do I Go About Finding Little Girls In Games. Examples [redacted]

This swift detection by AI allows human moderators to properly focus on their critical role — gathering information about the flagged content, substantiating its potential harm, and escalating the issue to law enforcement and other authorities when necessary.

Without the assistance of AI, expecting human moderators to address these tasks in a timely manner becomes near impossible, considering the sheer volume of content being generated on platforms every second. AI not only significantly accelerates the moderation process but also enables human moderators to concentrate on what they do best: complex tasks that require human empathy, understanding, and judgment.

> **If we are going to ask human beings to do this difficult work, we have a moral duty to provide them with working conditions that truly support their wellbeing and mental health.**
>
> 
>
> **Sharon Fisher**
> Global Head of Trust & Safety at Keywords Studios.
> - No more excuses

It's clear that, in today's online climate, human moderators alone cannot be responsible for moderating user-generated content. The toll on their mental health is too high, and with the amount of content generated daily, the potential to miss real-life threats is too dangerous. It is imperative that platforms employ technology to assist moderators in their important work.

In the upcoming section, we will explore both the advantages and the challenges that come with integrating AI into moderation processes.

ARTIFICIAL INTELLIGENCE

Considerations and Pitfalls in AI-driven Content Moderation.

Today, artificial intelligence algorithms and machine learning play important roles in moderating online environments. In fact, due to the sheer number of users online at any given time, AI is essential for effective moderation at scale. But while AI has the power to interpret interactions between players, it's not omniscient and relies on accurate, culturally aware, and platform-specific training in order to be effective.

Research, Training, Refinement

Effective training is essential for effective moderation. But therein lies another challenge: training the machine learning model specifically for any given online social environment.

Many machine learning models gather data from dozens to thousands of data sources, many from the public domain, while some privately developed tools are trained on anonymized data. This helps the model recognize a wide range of speech patterns, vocabularies, syntax, and cultural or subcultural nuances. But AI-driven moderation solutions are not one-size-fits-all. An ML model trained to look for violent language would be setting off all the alarm bells in the context of a first-person shooter game. So how do content moderation teams ensure the right things are being flagged?

Gaming-Centric Data

While general data can establish a strong foundation for AI moderation, gaming environments have their own ecosystems and niche communities, complete with specialized colloquialisms and game-specific player interactions. In order to properly train an algorithm for a niche community such as gaming, appropriate data must be collected, labeled, and used in the system's development.

> **Environment:** Gamers often find themselves in high-stakes, emotionally charged situations unique to the gaming experience. AI moderation tools need to understand the context within which gaming dialogue occurs to distinguish between playful competitiveness and bullying.
>
> **Lingo:** Gaming slang can have meanings vastly different from standard usage. AI models trained on data that does not represent this specialized language may misinterpret innocent banter as toxic behavior or hate speech.

To effectively moderate gaming interactions, it's crucial for your AI moderation tool to be well-versed in these specific nuances and lingo unique to gaming culture. By training on gaming-specific data, AI moderation systems can achieve the level of discernment needed to accurately interpret and moderate player interactions. When paired with an effective moderation strategy, it allows your moderation team to focus their time and energy on the issues that need a more human touch. Using a tool built on non-gaming data can increase the risk of false positives and false negatives in detection.

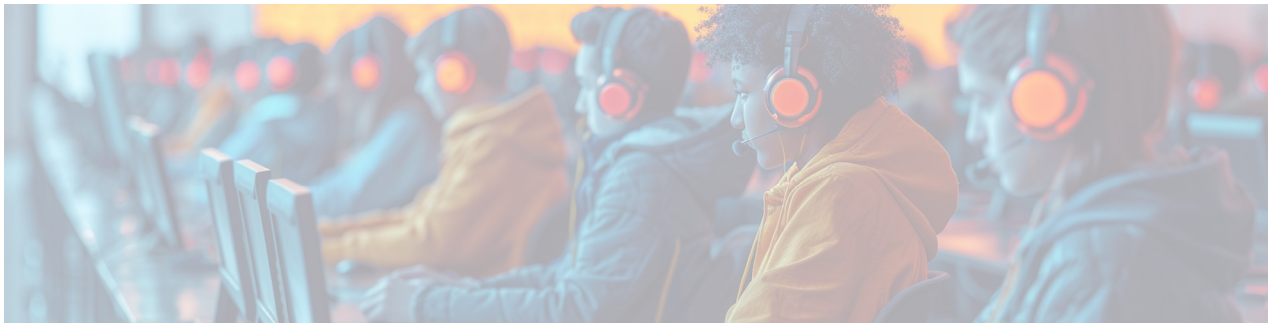
False Positives & False Negatives

Poorly trained and under-refined AI moderation algorithms can sometimes flag or remove content that isn't actually in violation of community guidelines. For instance, harmless jokes, friendly trash talk, sarcasm, or benign discussions on current events might be incorrectly identified as harmful. The model may also misunderstand the full context of a conversation, causing it to misinterpret statements. For example, it might flag a discussion about a controversial topic as offensive when it's actually a civil debate. Moreover, AI models may struggle to understand cultural nuances and context, leading to false positives when dealing with content that includes slang, idioms, or regional humor.

Conversely, AI systems may miss genuinely harmful content, including hate speech, harassment, or graphic material. This can lead to a toxic environment if such content goes unchecked. For example, if a text detection algorithm is only tuned to flag specific spellings of toxic or harmful words, bad actors may be able to evade moderation efforts by switching a letter for a number or intentionally misspelling keywords to avoid detection.

Of course, false positives and false negatives are both possible with humans too. The risk in using only AI moderation lies in the potential for systemic false positives/negatives. Where humans might miss one particular piece of context, mishear or misunderstand something, or otherwise just be a bit inconsistent (as is human nature), it's rare for human moderators to, at the scale possible with AI, misunderstand a whole class of harm. So AI tools have both the blessing and curse of consistency.

If your AI system doesn't properly understand the nuances of trash talk, then it will analyze every single instance of trash talk incorrectly. The flip side is that, where there's some risk of human moderators alone struggling to consistently moderate each user, AI systems give a strong guarantee that every user interaction is being analyzed exactly identically – for better or worse.



Accountability and the Right to Appeal

Users may often find themselves at odds with AI decisions. Due to its impersonal nature, they may also feel unheard or misunderstood. When users appeal, they want a human understanding of their situation that takes other signals like context, intent, and individuality into consideration — elements that AI cannot fully grasp. This can create an environment of distrust between your team and your players unless you have a robust and transparent appeal process.

Resource Allocation

Research, development, and maintenance of a “homegrown” AI moderation tool is no small investment. In addition to computational power needed, implementing AI moderation systems at scale also requires extensive and compliant data storage solutions. Developers and engineers must also continually fine-tune AI systems to respond to ever-changing online behaviors and language, including pop culture trends, game lingo, language nuances, and more. These costs are manageable for the largest platforms or for third-party service providers hosting models at scale, but can be prohibitive for mid-size studios considering homegrown AI tools.

Moderating Voice Chats

Voice chat has taken gaming to a whole new level, adding a more-human layer to play and fostering a greater sense of community across the globe. While AI-driven tools to transcribe speech into text are not new, most voice recognition tools are designed for the speaker to directly address the AI with a goal of being understood by the program. In the case of voice moderation for gaming however, players are speaking with each other with unique shorthand, heightened emotion, and highly variable audio quality. Monitoring and moderating voice chat conversations to detect toxic behavior like bullying, hate speech, and harassment requires voice-native detection, for a few key reasons:

1. **Emotional Nuance:** Many moderation teams rely on automatic speech-to-text transcription, but vital nuances can be lost when voice is simply transcribed into text. Transcription-based moderation misses out on these key factors like differentiating sarcasm or friendly trash talk from genuine harm.
2. **Demographics:** The context and impact of spoken words can vary significantly with the speaker's and listener's age, accent, gender presentation, and other demographic information which is evident only to tools that can understand tone.
3. **Responses:** The way other players react to a statement — with a laugh, an outcry, or with silence — is essential to understanding if something said was toxic or playful. This level of contextual understanding of a voice interaction often gets lost without direct voice native, or voice first moderation.

Furthermore, most voice recognition tools are designed for the speaker to directly address the AI with a goal of being understood (ie, speech-to-text transcription tools). In the case of moderation for gaming however, players are speaking with each other with unique shorthand, heightened emotion, and highly variable audio quality. Without a content moderation tool specifically designed to listen to and accurately understand the substance of a conversation, key

Voice chat needs to be moderated, too

In a study conducted by Modulate and Take This, researchers found that:

- 1 in 4 players perpetrated at least one incidence of offensive language in voice chat
- Among all players, 5.03% were flagged with at least one severe offense over the last 30 days
- Racial/cultural hate speech was the most common offense, constituting more than half of all offenses by all users by category

ARTIFICIAL INTELLIGENCE

Benefits of AI-driven Content Moderation.

Supplementing Traditional Moderation Tactics

In the evolving landscape of online gaming, AI moderation is a vital tool in managing community interactions — and scalability isn't the only benefit. Traditionally, player-generated reports are one of the (if not the only) main tools moderation teams use to identify and address toxic behavior in multiplayer games. Everyone's first reaction to online abuse is to report it, right?

> **28%** of respondents ages 10 to 17 filed a report to a game studio when they encountered harassment or hate in their game.

Unfortunately, the ADL's 2023 survey of gamers ages 10 to 45 shows that only 28% of respondents ages 10 to 17 filed a report to a game studio when they encountered harassment or hate in their game. 38% of adult respondents said they'd filed a report detailing toxic behavior. Reasons for the low rate of player-generated reporting include feeling the reporting process was too much work, feeling the behavior in question didn't meet the threshold to initiate a report, and feeling that the negative behavior is simply part of the gaming experience. Modulate has found in studies with its customer base that as few as 10-15% of reports are actionable, especially in competitive games where players have an unfortunate incentive to file reports against other players who just beat them and knocked them down the ranking ladder.

While player-generated reports alone are not effective in detecting and responding to toxicity, the addition of AI tools, especially those that are proactive, can improve the rates of response to toxic behaviors by removing the burden of reporting from the individual player.

With AI moderation tools to supplement player reporting, moderation teams can catch more toxic behavior, earlier — even when the bad actors are taking precautions to avoid detection.

Moderate at Scale

Let's say your game has 10 million monthly active users generating 5–20 million monthly hours of voice chat data. Within this data, there could be an estimated 7–25 million instances of harmful behavior — an unmanageable amount for human moderation teams, even at the largest studios. AI is able to process large volumes of data at an incredible speed. It's indispensable for effective moderation for large games. And the sooner you integrate an AI moderation tool into your moderation workflow, the easier that scaling will be.

Operate with Speed and Efficiency

As gaming communities grow, manual moderation becomes increasingly challenging, if not impossible. AI-driven systems can efficiently handle a large volume of user-generated content, making it scalable for games with small up to massive player bases. AI moderation allows some moderation actions to be taken immediately or empowers moderators to take immediate action (instead of hours or even days later once a moderator manually reviews a player report), which helps to prevent escalation of a negative interaction between players.

Get Data-driven Insights

AI moderation systems can generate valuable data and insights. Developers can analyze this data to understand patterns of toxic behavior, implement preventive measures, and continuously improve the overall gaming experience.

Build Trust

With a prioritized queue of the most critical issues and an AI tool to moderate at scale, moderation teams will spend a lot less time on minor issues or interpreting incomplete player reports. The result? Moderators free up significant bandwidth to more rapidly respond to infractions. This quick actioning helps nurture players' trust that a studio is taking toxic behavior seriously.

With the help of an AI tool to help human moderator teams to prioritize review items, moderators can be spared the time and mental bandwidth of having to dig through problematic and harmful content.

AI + HI

The Solution: A Hybrid Approach.

AI detection models are continually advancing. With the improvement of machine learning models, better training data, and a deeper integration of cultural and contextual understanding, AI moderation can become more adept at distinguishing between harmless and harmful content — but it will never be perfect.

A robust appeal system, ongoing refinement from a diverse team, strong community involvement, and a healthy environment for studios' moderators will be essential to mitigating the risks of both human and AI moderation. AI can provide scalability and efficiency, while moderation teams provide nuanced understanding, context, and cultural knowledge.

It takes time and money to implement an AI moderation system and adjust your human moderation processes, especially for large games with millions of players. But the reward — reduced player churn and reduced employee turnover — is worth the investment.

Content Moderation Roadmap:

An effective content moderation strategy within online gaming communities should follow a structured approach. A helpful framework to consider is Modulate's ToxMod voice moderation roadmap: triage, analysis, and escalation.

Triage: Initial Detection and Screening

While most text moderation skips this initial step (typically text moderation is cost-efficient enough to not require initial triaging phase), AI models assessing more complex forms of communication like voice, video and image will start here.

For cost-effective, comprehensive moderation of voice or video, a robust triage stage is required to quickly pinpoint which clips or frames are most likely to contain harmful content. Such a triage layer would avoid heavy-duty analysis – instead of analyzing all data, triaging first allows you to focus on dedicating resources to analyzing the highest risk pieces of content first. In the case of voice moderation, a strong triage approach might focus on things like expressions of anger, distress, or aggression; sudden changes in the engagement of certain participants; or demographic risk indicators.

Analyze: Contextual Understanding

The next layer of analysis delves deeper, considering the broader context of conversations. It evaluates nuance: slang, cultural references, and the specific history between players. With this information, the AI can better understand the true nature and severity of interactions and determine the next course of action.

Escalate: Human Intervention

After the AI has filtered out benign interactions and identified potential issues, the system escalates the most serious or complex cases to human moderators. These experts review the context provided by AI analyses to make informed, fair decisions and take proportional disciplinary action.

The combination of AI and HI is transforming the way gaming studios safeguard their communities and their superhero moderators. This hybrid approach leverages the best of both worlds, combining the scalability of AI with the nuanced understanding of human moderators.

A Collaborative Approach

The fusion of AI and HI into a hybrid model creates a balanced, thorough, and effective moderation solution with several advantages.

Thorough moderation: AI combs through large amounts of data to ensure all potential issues are flagged while HI addresses complex cases that require empathy and deeper cultural understanding.

Efficient resource allocation: AI handles high-volume, clear-cut cases, freeing up human moderators to focus on more ambiguous or sensitive content.

Enhanced accuracy: While AI offers speed, humans contribute contextual understanding, ensuring appropriate disciplinary action is levied out to the bad actors, not innocent players.

Swift responses: AI's rapid flagging combined with human verification accelerates the response to Real-Life Threats, thus preventing further harm, proactively addressing toxicity — even saving lives.

Protection from harm: AI identifies and removes obviously harmful and egregious content, protecting players and superhero moderators from exposure to psychologically damaging content.

Implementing an Effective AI + HI Strategy

No two studios have the exact same moderation needs. To capitalize on the strengths of AI and HI, you need an implementation plan tailored to your team and gaming environment. There are several key factors to consider:

1. **The amount of data you're handling:** The more data you have, the more you'll have to rely on AI to sift through the bulk of it and prioritize what matters.
2. **Gaming environment:** A first-person shooter will have a very different environment than a family-friendly game, and therefore very different expectations for language and banter.
3. **Common issues:** If your gaming community is frequently plagued with the same types of issues, you will want to train your AI accordingly and, if appropriate, automate moderation responses.
4. **Demographics:** The more diverse your user base, the more cultural and contextual awareness and training your AI and your HI will need. For example, while an AI moderation system may be multilingual, a studio will also need human moderators who have linguistic and cultural knowledge to understand those instances of infractions.
5. **Media Type(s):** There are several AI moderation tools on the market – some specialize in text and image detection like ActiveOS by ActiveFence, while other tools like Modulate's ToxMod focus on voice moderation.
6. **How you will recruit and train a moderation team:** You can hire internally or leverage the expertise of a proven and experienced vendor like Keywords Studios.
7. **Cost-benefit analysis:** Keeping all these things in mind, you should evaluate the costs of both AI tools and your personnel and the potential ROI. Your moderation solution should strike a balance between maximizing efficiency without compromising user experience.

Integrating AI and HI is not a one-and-done deal. It's an iterative process that should evolve with the changing dynamics of your community, advances in technology, and more diverse data.

CASE STUDY

Making Among Us VR Safer and More Inclusive.

The integration of AI-driven content moderation has proven to be a game-changer for Among Us VR developers Innersloth, Schell Games, and Robot Teddy, in creating a more positive player experience.

Prior to launching Among Us VR, Alexis Miller, Director of Product Management at Schell Games, recognized the critical role player communication played in a social multiplayer game like Among Us VR. Considering the original Among Us game appealed so heavily to younger audiences, the need for effective content moderation in Among Us VR was especially clear.

The Challenge

First, the Schell Games team needed to understand the usage rates of in-game communication. In Among Us VR, players can use Quick Chat to share pre-set messages like “hello,” “I’m ready,” “okay.” Players can also use Voice Chat to speak with one another. Early on, it was clear the most used communication channel for players was Voice Chat, which 90% of players use, while only 10% use Quick Chat. Acknowledging the need for a moderation system to monitor in-game voice conversations, Miller sought a solution that would maintain a safe and inclusive gaming environment. It became evident that relying solely on player reports was inadequate, not only for the potential of inaccuracies in player-generated reports, but also due to the high volume and time needed to review.

Simply asking moderators to listen to audio data to determine the best course of action was also impractical. “Having moderators review 100% of audio for offensive language was not a feasible solution because of the sheer volume of audio data,” says Miller. Schell Games also considered design decisions that could potentially reduce the harm of voice chat, such as only allowing in-game voice chat in private matches. That solution was not chosen because so few matches were truly private. “We realized that private matches often still include strangers because private match codes are shared on social platforms like Discord,” explains Miller.

In looking at the larger picture, Schell Games also recognized that while voice chat has problems related to toxicity and player safety, it also is a very popular feature and unique selling proposition for the VR version of the game.

Implementing AI Content Moderation

Understanding the limitations of a human-only approach to content moderation, the Schell Games team sought out an AI-driven solution to better protect their player base.

Soon after the launch of Among Us VR in November 2022, Schell Games implemented ToxMod, an AI-driven content moderation tool developed by Modulate for in-game voice chats, and partnered with Keywords Studios to bring trained, professional superhero moderators into the Schell Games moderation workflow.

Now armed with a tool that allowed expert moderators to efficiently address toxic user-generated content at scale and better understand player behavior through data-driven insights, the Schell Games team saw interesting data and tangible results.

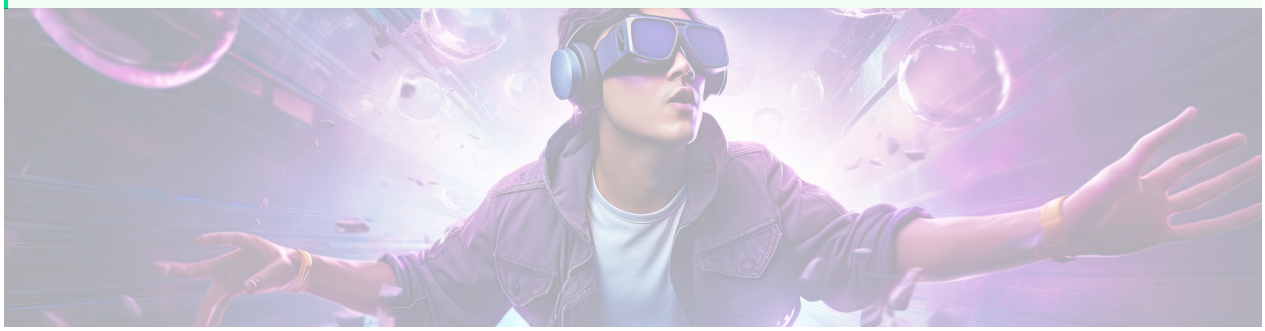
Laura Hall, Senior Player Support Specialist at Schell Games, estimates that at **nearly 50,000 hours of audio data per month**, the Schell Games moderation team would need to hire **more than four times their current staffing** if the studio did not use an AI tool like ToxMod.

The power of AI + HI: Key stats revealed

2-3K problematic players moderated per week, with an average of 30 players reviewed per hour.

80% of all moderator actions are in response to racial or cultural hate speech. Other infraction areas include sexual or gender hate speech and bullying.

49K+ hours of audio data per month processed by ToxMod and the moderation team.



Hall also emphasizes the continual refinement of ToxMod through human oversight, saying, “Constant review and modifications to the way our human moderators use ToxMod has resulted in a reliable tool that allows us to monitor and remove the most toxic utterances in the game.”

“Without ToxMod, we couldn’t moderate the community as efficiently as we do with the help of AI,” says a moderator on the project. “It pinpoints the important stuff that needs our attention.”

Working directly with Modulate to refine and train the AI model to the Among Us VR player community has allowed the detection accuracy of ToxMod to detect severe infractions with 95% accuracy, with accuracy increasing based on the severity of the infraction.

“It has gotten so accurate we are now looking forward to the benefits of automation within our processes to become more efficient,” says Hall.

“We still have human moderators to review ban appeals and to review more nuanced infractions,” she continues, “But after putting in the time to train the AI model, we can rely on it more and more to remove the most obvious toxicity from the game.”

Schell Games, with the integration of AI-driven content moderation, witnessed a paradigm shift in player support. The collaboration with Modulate and Keywords Studios to bring AI and HI into the content moderation ecosystem has paved the way for more efficient and scalable content moderation practices. As studios continue to invest in AI-driven solutions, the future of player support promises to be both technologically advanced and player-centric.

> **We couldn’t do what we’re doing without the combination of artificial intelligence (AI) and human intelligence (HI). AI allows us to identify potential toxicity in our game at scale, and HI ensures we’re doing the right thing about it.**

Laura Hall
Senior Player Support Specialist
at Schell Games

A Game-Changing Strategy for Safer Online Communities.

The indispensable role of content moderation has come to the foreground in recent years as more and more interpersonal interactions, community building, and connecting takes place online. The imperative to maintain a safe, inclusive, and thriving digital environment has been underscored by both societal expectations and legislative frameworks like the EU's Digital Services Act and the UK Online Safety Act. Furthermore, the business case for effective content moderation, as evidenced by the positive impact on player engagement and revenue, solidifies its central importance in the gaming industry.

The challenge in achieving a healthy virtual ecosystem lies in determining the most effective moderation approach to prevent the spread of misinformation, harmful content, hate speech and other forms of toxicity online. Rather than pursue an either-or approach, a hybrid model integrating Artificial Intelligence (AI) and Human Intelligence (HI), or AI + HI allows platform owners to **make accurate and nuanced moderation decisions at scale**.

Importantly, the AI + HI model also works to shield digital first responders from unneeded exposure to harmful content, thereby preventing potential long-term psychological harm.

AI forms a critical initial layer of content triage and analysis, effectively handling large volumes of data and surfacing potential issues. However, it is the nuanced understanding, cultural awareness, and context provided by superhero human moderators that become paramount in the escalate stage, especially in complex situations where AI may not have that context.

The implementation of AI moderation tools has proven essential in addressing challenges unique to gaming environments. From supplementing player reporting and moderating at scale to focusing on community building and safeguarding moderators, the benefits of AI in gaming moderation are evident.

The risk of relying solely on AI tools for moderation may lead to false positives and negatives, lack of context understanding, biases, and the need for a robust appeal process. Thus, the integration of AI into a content moderation solution should be viewed as an ongoing process, constantly refined based on user feedback, diverse data sources, and evolving language trends. Human moderators' ability to discern sarcasm, navigate linguistic trends, address complex situations, and escalate issues of illegal content remains unmatched.

The case for the hybrid model, AI + HI, is not just about moderating content; it's about enhancing the overall user experience, reducing churn, and fostering healthier communities.

The goal is clear — to leverage technology for efficiency without compromising the profound understanding and empathy that superhero human moderators bring to the digital realm. The result is not just effective moderation; it's the creation of digital spaces that are safe and truly enjoyable for every user.

Keywords Studios Imagine More | MODULATE | ActiveFence | Take this.

About the Gaming Safety Coalition:

Launched in 2024, the Gaming Safety Coalition represents a strategic alliance among leading gaming Trust & Safety entities dedicated to creating more robust, safer, and more resilient gaming environments. The coalition embodies its partners' shared commitment to improving player and moderator well-being within gaming communities. The Coalition's founding organizations include Keywords Studios, Modulate, ActiveFence, and Take This.

Learn more