Keywords Studios
Imagine More

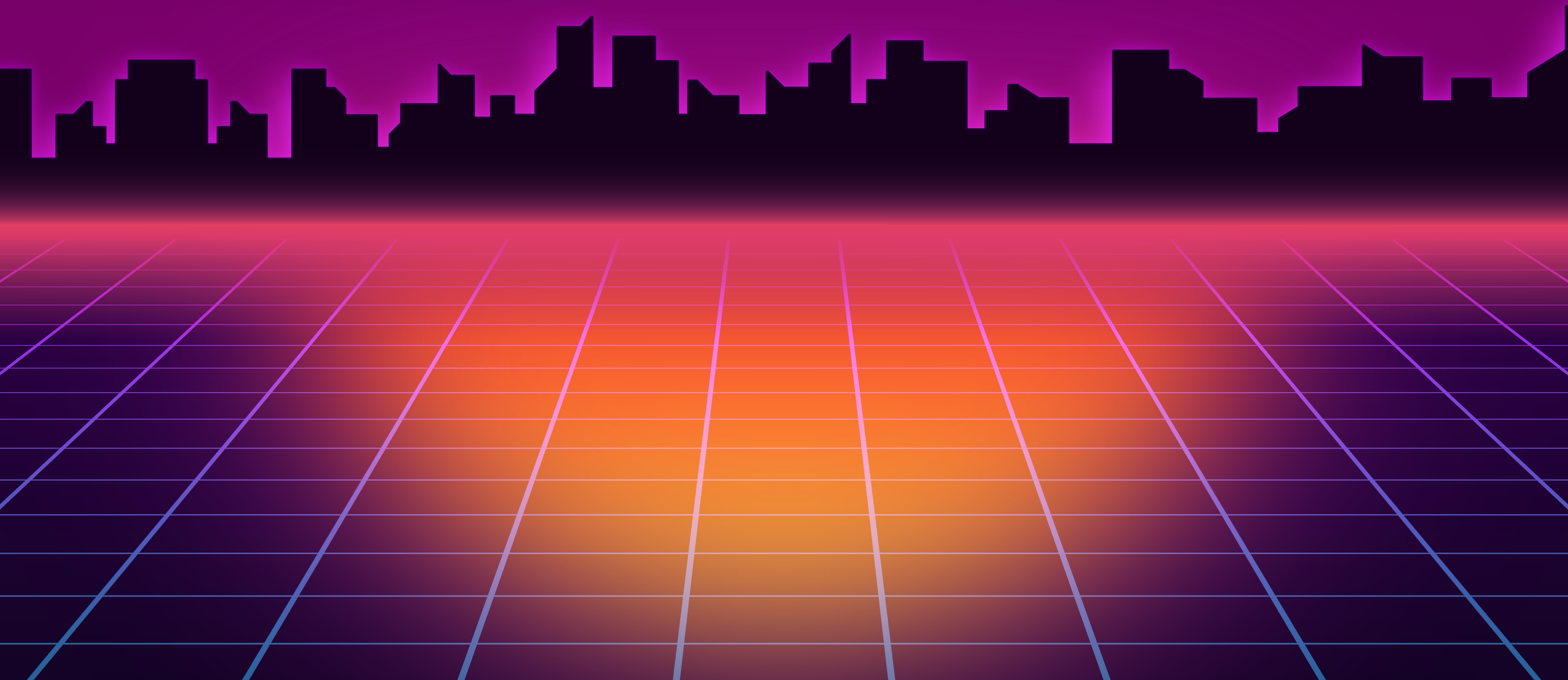
MODULATE

ActiveFence

Take this.

GAMING SAFETY Coalition

Tipsheet



Gaming Safety Coalition

TIPSHEET

Content Moderation Best Practices.

Content moderators serve on the front lines of online safety to detect and combat problematic content and behaviors. As games continue to become more akin to social platforms, it is essential for trust and safety teams to adopt moderation best practices that prioritize both user safety and the well-being of “superhero” first responders.

The Gaming Safety Coalition is dedicated to sharing content moderation best practices, with the goal of improving online experiences for players and moderators. This tip sheet outlines key strategies for effective content moderation, encompassing approaches to building multimodal safety strategies, protecting superhero moderators, and planning for ongoing improvements, especially when using AI tools to bolster the scale of moderation and safety practices.

Build Multimodal Safety Strategies

The strongest safety strategies are platform-wide and cover all available modes of communication, from voice chat to image sharing and in-game player interactions. Successful trust and safety teams will implement foundational moderation practices that are built into a game’s design and feel engaging and intuitive for players.

Prioritize User Engagement with Safety Features: Consider integrating safety features in a way that engages users and makes safety measures more interactive and immersive. This approach can include gamification, educational content, and incentives for positive behavior. By making safety a more engaging aspect of the platform, users are more likely to interact with it and adhere to safety guidelines.

Adopt a Proactive Approach to Safety by Design: Embrace a safety-by-design approach that considers potential risks and safety measures early in the development process. A consultation with a Trust & Safety expert or a community lead will save you time, resources, and money. This proactive approach involves conducting risk assessments, implementing safety controls like in-game reporting tools, and continuously monitoring and updating safety features. By embedding safety considerations into the core design and functionality of the platform, potential safety issues can be addressed more effectively.

Utilize Off-Platform Intelligence and Collaboration: Recognize the importance of leveraging off-platform intelligence and collaborating with external entities, including regulatory bodies, law enforcement, and industry partners. By staying informed about threats and trends in the broader ecosystem, platforms can better anticipate and mitigate potential safety risks. Additionally, collaborating with external stakeholders (while still prioritizing user privacy) allows for the sharing of threat intelligence and best practices, enhancing overall safety efforts, and potentially creating a positive impact in the “real world”.

Protect Your Superhero Moderators

Moderators are often the first to be exposed to problematic content in the online safety workflow. How can trust and safety teams build more resilient teams that not only protect players but also account for the longevity and mental health of the moderators themselves?

It may seem obvious in 2024, but experience tells us otherwise: When assembling a moderation team, please refrain from appointing your most dedicated and active players to these roles as unpaid and untrained moderators. Moderation is too critical a task to be left to volunteers, especially when it comes to harmful content.

Prioritize Thoughtful Recruitment Practices: Recruit, interview, and onboard moderators based not only on their language skills but also on their resilience, communication abilities, empathy, and bias awareness. Craft job descriptions and evaluation tests that assess candidates' suitability for the challenging content they may encounter. This approach ensures that moderators are well-equipped to handle their responsibilities while prioritizing their mental health and wellbeing. It's crucial that we recognize and champion the diverse skill set that moderators bring to a challenging role.

Establish Accessible Wellbeing and Resilience Programs: Implement proactive wellbeing programs that provide moderators with accessible and trustworthy mental health resources and support on a daily basis. Leadership should offer continuous support from the top down, making mental health resources available during and outside of work hours. By focusing on prevention rather than reactive measures, platforms can better support moderators in managing the potential for mental distress caused by their work.

Foster Empathetic and Compassionate Leadership: Cultivate leadership that views moderators as more than just performance metrics or numbers on a spreadsheet. Leaders should recognize moderators as true "superheroes" and demonstrate empathy and compassion at every level of the organization. This acknowledgment means providing moderators with a complete set of tools, training, and resources. These should aim at not only safeguarding their mental well-being but also supporting their career development and success.

Plan for Ongoing AI Improvements

AI safety tools have grown in use in recent years and play a crucial role in automating and enhancing content moderation processes, enabling platforms to swiftly identify and address problematic content at scale. However, as AI technology continues to advance, it is imperative for platforms to plan for ongoing improvements to ensure the fairness, effectiveness, and ethical integrity of these AI tools.

Ensure Fairness and Equity in Model Design: Place great emphasis on ensuring the fairness of internal model design and performance. Evaluate whether AI models perform equally across different demographic axes by using balanced in-domain datasets. Implementing this approach helps to mitigate potential biases and ensures that the AI models are equitable across diverse user groups.

Stay Abreast of Cutting-Edge Research on Bias Mitigation: Actively monitor cutting-edge methods for uncovering and mitigating model biases. Explore research papers and initiatives focused on bias amplification, bias detection, and debiasing techniques in AI models. Seek and incorporate moderator feedback around false negatives and false positives on an ongoing basis. Consider implementing novel post-processing methods and leveraging techniques from other domains, such as language interpretation, to mitigate biases effectively in AI models used for content moderation.

Promote Collaboration and Improve Access to Data: Foster collaboration and discussion within the AI research community by organizing workshops and events focused on ethical AI development. Provide access to extensive datasets, particularly valuable in the audio research domain, to encourage further exploration and research into ethical and effective machine learning models. By improving access to data and promoting collaboration, the development of ethical AI models for content moderation can be accelerated, benefiting society as a whole.

Get in touch with the Gaming Safety Coalition to learn more about player safety and content moderation: hello@gamingsafetycoalition.com

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