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

GAMING SAFETY Coalition

5 MYTHS IN TRUST & SAFETY FOR GAMING

Gaming Safety Coalition

TIPSHEET

Debunking Common Myths in Trust & Safety for Games.

Trust and safety in gaming isn't just about removing bad actors — it's about actively building better communities. But as gaming platforms evolve, new legislation emerges, and AI tools grow more sophisticated, keeping players safe has never been more complex.

To help gaming studios stay ahead of the curve, we're busting five outdated myths and replacing them with grounded, practical truths from today's leading voices in gaming safety.

Myth #1: Moderation is censorship, and it's easy to tell what content should be taken down and what should stay up.

Truth: Moderation empowers gaming studios to keep users safe, without censoring content.

Gaming can be edgy, violent, and even explicit — and that's okay. The goal of moderation isn't to change the nature of gaming — it's to keep players safe. Context matters. What's acceptable in Call of Duty may be totally off-base in Animal Crossing.

To build effective moderation policies, consider:

- What's always unacceptable? Hate speech, slurs, and threats rarely depend on context.
- Who is your community? Age, gender, region, and culture all affect expectations.

What kind of community are you building? Let your code of conduct reflect your values — and enforce accordingly.

> **Moderation itself doesn't equate to censorship. That's like saying that having police makes you a police state. It's more important how you use moderation.**
>
> **— Tomer Poran, ActiveFence**

Myth #2: Content moderation is only a concern for large social media platforms.

Truth: Games are social platforms — and moderation is key to sustainable growth.

Games are social spaces. And like any community, how you moderate them influences everything from player retention to revenue. In fact, **research shows** that players are less likely to stay in or spend money on toxic gaming communities, even if the player hasn't personally been targeted.

Key considerations for game-specific moderation:

- Ephemeral content matters. Voice and in-match chat may not be recorded or visible after a session, but they still affect player experience.
- Real-time moderation isn't always necessary — but customization is. Build a strategy that fits your game's pace.
- Smaller communities benefit the most from early action. How you moderate content directly shapes your community's sustainability, safety, and success.

> **Smaller communities are much easier to shape norms in. Starting moderation early helps instill healthy norms in the community as it grows larger.**
>
> **— Dr. Elizabeth Kilmer, Take This**

Myth #3: AI technology can completely replace human content moderators.

Truth: AI tools work best when paired with human validation and oversight.

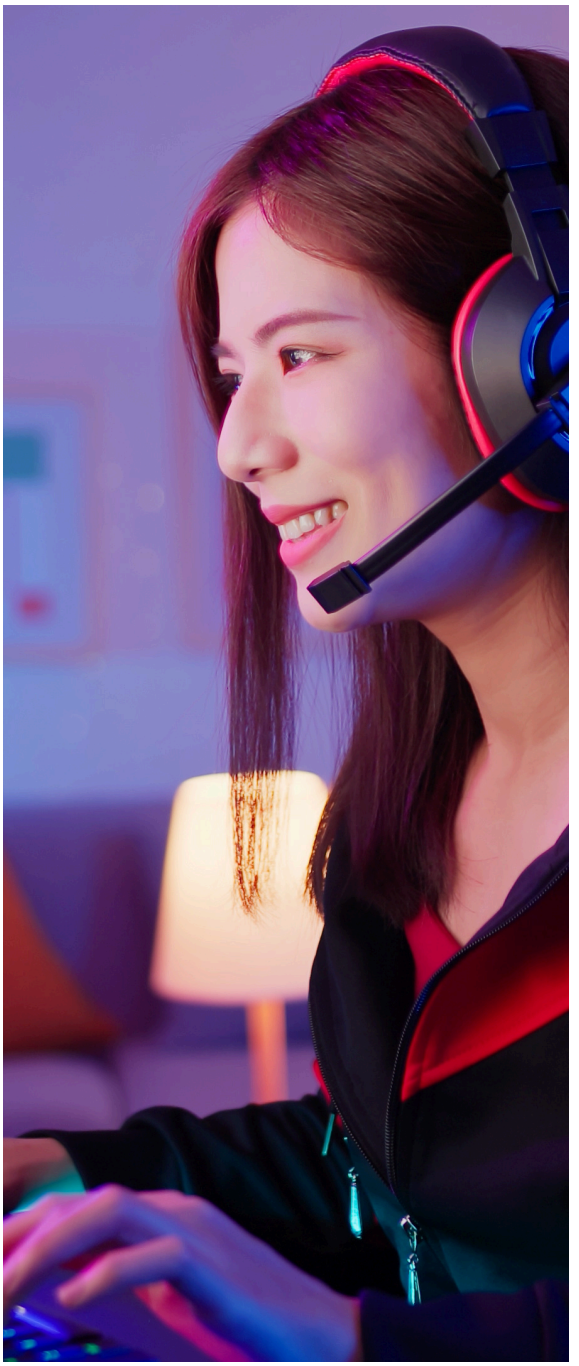
Only **25% of harms get reported** in gaming communities, and the most insidious don't get reported at all. AI helps scale moderation — identifying harmful content that would otherwise fly under the radar. But it can't understand nuance, tone, or slang like humans can. Human moderators provide the necessary context to validate AI results and address complex edge cases.

Best practices:

- Use AI to flag and filter at scale.
- Pair AI with trained human moderators who can interpret context.
- Even a small human moderation team makes a huge difference in validating automated decisions.

> **AI can address the problem of scale. For example, it would be impossible to manually moderate every user interaction in Fortnite. However, overreliance on a nonspecific AI tool can create pitfalls.**
>
> **— Mark Nolan, Modulate**

Myth #4: Content moderators are biased and remove content based on personal belief.



Truth: Truth: While no one is without bias, skilled moderators are trained to follow policy — not personal beliefs.

To support consistency and fairness, gaming studios should:

- Pre-test moderators to ensure awareness of bias and openness to diverse perspectives.
- Tie moderation decisions to clear, game-specific policies.
- Run regular calibration sessions to check for selective enforcement and address discrepancies.

Provide ongoing training and mental health support to help moderators manage stress and stay sharp.

> **Moderators are not merely button pushers; they are skilled professionals with language proficiency, analytical skills, swift decision-making, and current knowledge of pop culture, geopolitical changes, in-game language, and gameplay.**

— Sharon Fisher, Keywords Studios

Myth #5: Age verification is a privacy risk, not a safety tool.

Truth: The real privacy risk isn't age verification, it's allowing kids to access things that are meant for adults.

As technology advances, age verification doesn't need to be a privacy tradeoff. Modern age verification tools are:

- Device-based and privacy-first. They don't require sensitive data to be shared externally.
- Reusable. Once verified, users can confirm their age across platforms without repeating the process.
- Essential for safety. Without age assurance, younger players are left vulnerable to inappropriate content or interactions.

> **Just like amusement park rides have height checks, digital spaces should have age-appropriate zones that are safe, fun, and designed for who is using those spaces**
>
> **— Jeff Wu, k-ID**

Gaming should be bold, immersive, and fun — and that's only possible when players feel safe, respected, and empowered to connect authentically. By challenging outdated myths around moderation, AI, and community dynamics, today's studios have a unique opportunity to lead with intention.

Trust and Safety isn't about restricting creativity or building walls — it's about designing smarter systems that align with your game's values, meet your players where they are, and scale with integrity. Whether you're managing a fast-paced voice chat environment, building out age-appropriate spaces, or supporting a team of human moderators, the goal is the same: to shape digital spaces where everyone can play — and thrive.

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.