

Towards Ethical AI Moderation in Multiplayer Games

Participants engaged in guided design exercises, creating potential AI moderation systems using worksheets that prompted consideration of user stories, interfaces, and key features.



Participants viewed AI moderation as enforcing rules and responding to violations, but often unreliably. While some saw this as enhancing game safety, concerns were raised about **incompetence, lack of transparency, and potential overreach.**

AI moderation was perceived as a ruling entity, sometimes acting against community interests. Players worried about **top-down rule enforcement and hidden agendas**, while industry professionals debated **who should design and implement** these systems.

Participants likened AI moderation to a judge issuing punishments, with mixed views on whether its **lack of emotion** led to more or less **consistency** than human moderators. Many found the AI's reasoning and appeal processes **unclear**.

While valued for its **tireless** work in addressing toxicity, AI was seen as better suited to **assisting** human moderators rather than making high-level decisions **independently**.

Implement **transparent** AI moderation systems with clear explanations of decision-making processes.

Design **privacy-preserving** AI moderators that minimize unnecessary data collection.

Develop inclusive AI moderation tools that account for player diversity and **avoid bias.**

Create **hybrid** moderation systems that **combine** AI efficiency with human judgment for complex cases.

RQ2

How can AI moderation be implemented in multiplayer games in **responsible and ethical** ways?

