Explaining Toxicity in Multiplayer Games

Timothy Holland
The University of Melbourne
Parkville, VIC 3010
tholland1@student.unimelb.edu.au

Lucy Sparrow, Wally Smith
The University of Melbourne
Parkville, VIC 3010
lucy.sparrow@unimelb.edu.au, wsmith@unimelb.edu.au

Keywords
Multiplayer Games, Toxicity, Content Moderation, Artificial Intelligence

INTRODUCTION
A majority of multiplayer game players have reported experiencing or witnessing abuse while playing online (Anti-Defamation League 2020). Such "toxicity" manifests in various forms, ranging from fleeting verbal interactions like "trash-talking", the belittling of other players, to more strategic and behavioural acts such as cheating, the use of an exploit to gain an unfair advantage (Kowert 2020). The persistence of toxicity indicates that content moderation - the formal "governance mechanisms" deployed to "facilitate cooperation and prevent abuse" (Grimmelmann 2015, p.47) - is failing players in practice.

Moderation primarily takes the form of reporting tools (Kou and Gui, 2021) utilised by moderators (Seering et al. 2019), communities (Seering 2020), and increasingly algorithms (Gillespie 2020), resulting in warnings (Seering et al. 2019), content removal (Srinivasan et al. 2019), or account bans (Kou 2021). However, in practice, the normalisation of toxicity has led to the underutilisation of these tools (Beres et al. 2021). Additionally, their outcomes are often perceived as opaque and unfair, diminishing their effectiveness in reforming behaviour (Ma et al. 2023b).

The failure of moderation is a design problem. Research indicates that offending players often act in good faith, engaging in information-seeking behaviours on online forums to overcome the opaqueness of moderation outcomes (Kou and Gui 2020). Moderation systems that provide *explanations* alongside decisions are perceived as more transparent and fair, increasing the likelihood that offenders will understand how to modify their behaviour (Ma et al. 2023a). Nevertheless, such explanations remain overly simplistic, typically justifying decisions by reference to the chat log that was deemed toxic (see Figure 1).

A question naturally arises: could other forms of explanations that centre transparency improve the perceived fairness of moderation outcomes? Our work-in-progress research will answer this question by deploying Explainable Artificial Intelligence (XAI) to automate the generation of moderation explanations. In general, XAI research focuses on developing techniques to make the decision-making processes of automated systems clear and understandable to users (Gunning and Aha 2019). We intend to examine the use of XAI techniques to generate moderation explanations through a within-subjects experiment. In this experiment, participants

Proceedings of DiGRA Australia 2025

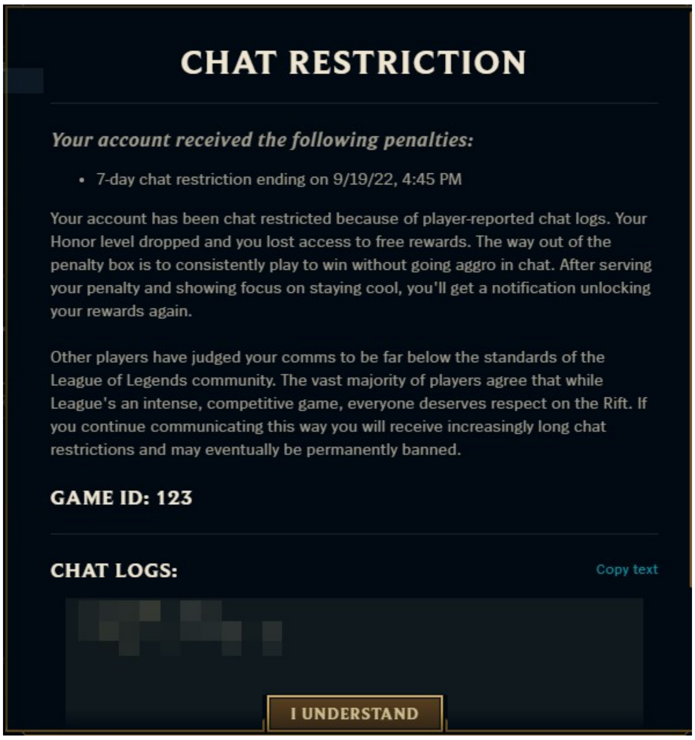

Figure 1: League of Legends Penalty Notification (League of Legends Support, 2022).

(adult players of multiplayer games) will engage with fictional moderation notifications justified by different forms of explanations and will self-report their perceptions across fairness measures and open-answer questions.

Although data collection hasn't occurred, our preliminary outcome focuses on our explanation design. Each explanation under examination is composed of two independent variables, an *explanation type* delivered by an *interaction style*. We have chosen these two variables as they pertain to open questions within the XAI literature. In particular, we have chosen three levels for each variable to stage a comparison between current practices, the literature's postured ideal, and a promising alternative.

Explanation type is the means by which decisions are justified (see Table 1 for examples). *Referential* explanations are used in practice, justifying moderation decisions by reference to toxic chat logs. *Counterfactual* explanations are held as an explanation ideal for decision-subjects as they don't expose the underlying logic of decision-making algorithms, thereby protecting trade secrets and preventing the gaming of outcomes (Wachter et al. 2018; Ribera and Lapedriza 2019). Counterfactuals justify a decision by way of external facts; what would need to change for the decision to be otherwise. For instance, a counterfactual for a loan application might indicate an additional $10,000 in income is required to qualify. Within our context, a counterfactual is a chat log edited to be non-toxic through word replacement or removal. However, generating coherent text-based counterfactuals is a difficult task, as the editing invariably affects the semantic content (Madsen et al. 2022). *Attribution* explanations present a promising alternative as they simplify the task from identifying what would be non-toxic to identifying what is toxic. These explanations highlight the degree to which each input feature contributes to the classification (Ribeiro et al. 2016; Lundberg and Lee 2017). In our context, this refers to the degree to which the algorithm considers a word to be toxic.

You are a [toxic-word]	You are a [non-toxic-word]	You are a [toxic-word]
(a) Referential	(b) Counterfactual	(c) Attribution

Table 1: Example Explanation Types

Interaction style is the mode by which explanations are delivered. *Static* interactions are used in practice, presenting explanations without further engagement. *Conversational* interactions have been heralded as an interaction ideal as people are cognitively wired to produce and consume explanations in a socially constructed way (Miller 2019; Ribera and Lapedriza 2019; Liao et al. 2020; Ehsan et al. 2024). A conversational agent, facilitated by a Large Language Model, will tailor underlying explanation types to user questions. This approach has shown promise in decision-support applications like medical diagnosis (Slack et al. 2023). However, questions remain over whether the same holds in the non-cooperative context of content moderation. *Explorative* interactions present a promising alternative by centering user control. These interactions expose the underlying classification system, allowing users to simulate alternative inputs and receive corresponding explanations (Bertrand et al. 2023).

At the conclusion of our study, we will have two unique contributions. For the XAI literature, we will provide an examination of how explanation type and interaction style combine to affect the perceptions of algorithmic decisions. For the content moderation literature, we will provide an evaluation of how different forms of explanations affect players' perceived fairness of automated moderation. More generally, our research will form part of a growing movement aimed at fostering legitimacy, accountability, and transparency within online governance, with the ultimate goal of promoting more inclusive online communities.

BIO

Timothy Holland is a Master of Computer Science student at the University of Melbourne. He previously wrote his Honours thesis on the links between colonialism and technological development, situating modern alienation from this lens. He plans to pursue a PhD at the intersection of Philosophy and Computer Science.

BIBLIOGRAPHY

Anti-Defamation League. 2020. "Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020." Anti-Defamation League. https://www.adl.org/resources/report/free-play-hate-harassment-and-positive-social-experience-online-games-2020.

Beres, Nicole A, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. "Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. CHI '21. New York: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445157.

Bertrand, Astrid, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. "On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. CHI '23. New York: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581314.

Ehsan, Upol, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. "The Who in XAI: How AI Background Shapes Perceptions of AI Explanations." In *Proceedings of the CHI Conference on Human Factors in*

Computing Systems, 1–32. CHI ’24. New York: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642474.

Gillespie, Tarleton. 2020. “Content Moderation, AI, and the Question of Scale.” *Big Data & Society* 7 (2): 1–5. https://doi.org/10.1177/2053951720943234.

Grimmelmann, James. 2015. “The Virtues of Moderation.” *Yale Journal of Law and Technology* 17:42–109.

Gunning, David, and David Aha. 2019. “DARPA’s Explainable Artificial Intelligence (XAI) Program.” *AI Magazine* 40 (2): 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Kou, Yubo. 2021. “Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community.” *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 334:1-334:21. https://doi.org/10.1145/3476075.

Kou, Yubo, and Xinning Gui. 2020. “Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation.” *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 102:1-102:27. https://doi.org/10.1145/3415173.

Kou, Yubo, and Xinning Gui. 2021. “Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. CHI ’21. New York: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445279.

Kowert, Rachel. 2020. “Dark Participation in Games.” *Frontiers in Psychology* 11 (November):598947. https://doi.org/10.3389/fpsyg.2020.598947.

League of Legends Support. 2022. “In-Client Penalty Notifications FAQ.” Riot. October 12, 2022. https://support-leagueoflegends.riotgames.com/hc/en-us/articles/205097293-In-Client-Penalty-Notifications-FAQ.

Liao, Q. Vera, Daniel Gruen, and Sarah Miller. 2020. “Questioning the AI: Informing Design Practices for Explainable AI User Experiences.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. CHI ’20. New York: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376590.

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77. NIPS’17. New York: Curran Associates Inc.

Ma, Renkai, Yao Li, and Yubo Kou. 2023. “Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. CHI ’23. New York: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581097.

Ma, Renkai, Yue You, Xinning Gui, and Yubo Kou. 2023. “How Do Users Experience Moderation?: A Systematic Literature Review.” *Proceedings of the ACM* *on Human-Computer Interaction* 7 (October):278:1-278:30. https://doi.org/10.1145/3610069.

Madsen, Andreas, Siva Reddy, and Sarath Chandar. 2022. "Post-Hoc Interpretability for Neural NLP: A Survey." *ACM Comput. Surv.* 55 (8): 155:1-155:42. https://doi.org/10.1145/3546577.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267 (February):1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD '16. New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Ribera, Mireia, and Agata Lapedriza. 2019. "Can We Do Better Explanations? A Proposal of User-Centered Explainable AI." *CEUR Workshop Proceedings*, March.

Seering, Joseph. 2020. "Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 107:1-107:28. https://doi.org/10.1145/3415178.

Seering, Joseph, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. "Moderator Engagement and Community Development in the Age of Algorithms." *New Media & Society* 21 (7): 1417–43. https://doi.org/10.1177/1461444818821316.

Slack, Dylan, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. "Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel." *Nature Machine Intelligence* 5 (8): 873–83. https://doi.org/10.1038/s42256-023-00692-8.

Srinivasan, Kumar Bhargav, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. "Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 163:1-163:21. https://doi.org/10.1145/3359265.

Wachter, S., B. Mittelstadt, and C. Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law and Technology* 31 (2).