Co-Designing AI Tools for Inclusive Online Environments

Ren Galwey, Dahlia Jovic & Mahli-Ann Butt
School of Culture and Communication
Faculty of Arts
University of Melbourne
ren.galwey@unimelb.edu.au; dahlia.jovic@unimelb.edu.au;
mahliann.butt@unimelb.edu.au

Yige Song, Sable Wang-Willis,
Lucy Sparrow & Eduardo Araujo Oliveira
School of Computing and Information Systems
Faculty of Engineering
University of Melbourne
yige.song1@unimelb.edu.au; sable.w@unimelb.edu.au;
lucy.sparrow@unimelb.edu.au; eduardo.oliveira@unimelb.edu.au

Keywords
Online harassment, gender-based harassment, online games, artificial intelligence, upstander, moderation

INTRODUCTION
Toxicity in online spaces, including gaming, has reached unprecedented levels, with gendered harassment in particular at an all-time high. In Australia, 65 per cent of girls and young women have reported being harassed or abused online – higher rates than the global average (Hermant 2020; Powell and Henry 2015). This harassment includes persistent unwanted contact via messaging, email and social media, location monitoring through social media posts, as well as bombarding victims with threatening language and imagery (Amnesty International 2018). In online games, harassment occurs both during gameplay and in game-related forums and communities (Fox and Tang 2017; Massanari 2017; Shaw 2012). The gaming industry and online platforms have responded with various moderation methods, including automated tools and human moderators, implemented to varying degrees of success (Gorwa et al. 2020; Yang et al. 2023). Recently, digital platforms have shifted towards artificial intelligence (AI) for moderating player communities and social media. While AI brings forth various advantages to efficiency and decreases the human emotional burden of moderation, it also raises ethical concerns about users' rights, autonomy, and privacy. Additionally, while most AI moderation approaches focus on punitive measures, there is a growing recognition of the need for more proactive and supportive tools that foster positive interactions and provide targeted support to vulnerable groups (Reid et al. 2022; Xiao et al. 2022). This paper will explore the findings of the GAIM (Gaming AI Moderation) project and its successor, the AI Ally project, offering insights into the potential of AI-assisted moderation and support systems for victims of harassment in gaming and other online spaces.

Our research reimagines AI's role in online communities. As part of the GAIM project, we conducted 26 in-depth interviews with players and industry professionals to gauge perceptions of AI moderation in multiplayer online games. We also analysed moderation practices in game-adjacent communities, such as Discord servers, and

Proceedings of DiGRA Australia 2025

held workshops to explore alternative roles for AI beyond punitive moderation (Sparrow et al. 2024). The findings from the GAIM project highlighted the limitations of punitive AI systems and identified a desire for more supportive, user-centred interventions, communicated through metaphors such as “teacher”, “gardener” and “upstander”.

While the GAIM project focused on moderation in gaming spaces, it highlighted the broader issues of online harassment and toxicity. Events like Gamergate exemplify how hegemonic ‘gamer’ culture can fuel harassment across social media and other digital platforms, further reinforcing the marginalisation of women and other vulnerable groups (Massanari, 2017). The AI Ally project examined gendered online harassment in this broader online context. We surveyed 230 girls, young women and gender-diverse individuals aged 14-25 to understand their experiences with online harassment and ‘upstander’ (Davidovic et al., 2023) intervention. Our research revealed a high prevalence of harassment, with 44% of our respondents “often” or “always” experiencing gendered harassment on at least one social media platform. 81% of respondents reported experiences of sexist comments, and 63% encountered sexual harassment. The most common responses to harassment included blocking harassers (74%) and ignoring the harassment (55%). Platform reporting mechanisms and legal responses were largely perceived as ineffective.

In instances of witnessing harassment, bystanders noted barriers to intervention (upstanding) on current online platforms, such as concerns for personal safety (52%) and fear of becoming targets (46%). Survey respondents were asked to rank the functions AI Ally could fulfil to help them overcome these barriers. The most preferred function was to document the harassment and draft a report or summary for the user. As such, we are developing the “TABBI” dashboard, which will assist victims of harassment in identifying the type of harassment they have experienced and prepare them for the next steps in various reporting mechanisms (such as eSafety). To ensure the dashboard's functionality is designed for realistic user scenarios, we have created user personas based on aggregate data and experiences from the survey respondents (Figure 1). This ensures that “TABBI” will respond appropriately to the *actual* needs of the community it serves.

Figure 1. AI Ally personas designed from survey responses.

The GAIM and AI Ally projects highlight the potential for AI to transform toxic online and gaming environments into safer, more inclusive spaces. By reimagining AI's role from a punitive enforcer to a supportive ally, we can foster communities of practice that encourage positive behaviours. AI systems have the potential to provide real-time support and protection for vulnerable community members. As our research progresses, we aim to develop and test AI-powered tools that embody the upstander – a role that our survey respondents identified as important in combating online harassment. We plan to collaborate with gaming and social media platforms to implement and evaluate these tools in live environments. It is crucial that these interventions empower users and that they are ethical, safe, and tailored to individual and community needs.

BIO

Ren is a law graduate (JD) and research assistant at the University of Melbourne and the University of Sydney. Their research interests include video games, intellectual property, artificial intelligence, data and privacy. They also run Research Rendered, an academic design and research translation service.

Dahlia completed her BA(Media&Comm)(Hons) at the University of Sydney and is a research assistant at the University of Melbourne. Her research interests include game studies, technology, artificial intelligence, and digital cultures.

Yige is a PhD student at the University of Melbourne, researching data science, AI and Education. His research interests also include social studies and psychology, with a focus on applying his work to societal improvement.

Sable is a Masters student in Software Engineering at the University of Melbourne. Her research interests include web development, game development, DevOps, and social justice.

Lucy is an Associate Lecturer in the Human-Computer Interaction Group at the University of Melbourne. Her award-winning research focuses on digital ethics, with a particular interest in multiplayer games and game design.

Eduardo is a Senior Lecturer in Software Engineering in the School of Computing and Information Systems at the University of Melbourne. His expertise is in the use of artificial intelligence to model and assist tertiary students in digital learning environments. Eduardo uses machine learning and natural language processing combined with models of self-regulated learning to guide his research.

Mahli-Ann is a Lecturer in Cultural Studies at the University of Melbourne. She is the co-curator of Feminine Play (2024), exhibition producer for Pride at Play (2023), and lead editor of the forthcoming anthology, *The Post-Gamer Turn*. Her research interests focus on questions of diversity in the digital cultures and creative industries of videogames.

ACKNOWLEDGMENTS

Special thanks to the Centre for Artificial Intelligence and Digital Ethics (CAIDE) at the University of Melbourne for financially supporting the GAIM project, and to the eSafety Commissioner for funding AI Ally. We also appreciate Girl Geek Academy for their invaluable partnership and our Advisory Board for their ongoing guidance. Thank you also to Maddy Weeks, our social media officer, for getting the word out about the project. Finally, a heartfelt thank you to our research participants—your insights have been instrumental in shaping our work and driving our mission for safer online spaces.

BIBLIOGRAPHY

Amnesty International (February 7, 2018). "Australia: Poll reveals alarming impact of online abuse against women." https://www.amnesty.org.au/australia-poll-reveals-alarming-impact-online-abuse-women.

Davidovic, A., Talbot, C., Hamilton-Giachritsis, C., and Joinson, A. (2023). "To intervene or not to intervene: Young adults' views on when and how to intervene in online harassment." *Journal of Computer-Mediated Communication*, 28(5), zmad027.

Fox, J., and W. Y. Tang. (2017). "Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies." *New Media & Society*, 19(8): 1290-1307. https://doi.org/10.1177/1461444816635778.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." *Big Data & Society*, 7(1). https://doi.org/10.1177/2053951719897945.

Hermant, N. (October 5, 2020). "Young Australian women cop more online harassment than global average, report finds." ABC News. https://www.abc.net.au/news/2020-10-05/young-australian-women-online-abuse-harassment-planinternational/12725286.

Massanari, A. (2017). "Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures." *New Media & Society* 19, no. 3 (2017): 329–46. https://doi.org/10.1177/1461444815608807.

Powell, A. & Henry, N. (2015). "Digital harassment and abuse of adult Australians: A summary report." https://www.parliament.nsw.gov.au/lcdocs/other/7351/Tabled%20Document%20-Digital%20Harassment%20and%20Abuse%20of%20A.pdf.

Reid, E., et al. (2022). "Feeling good and in control: In-game tools to support targets of toxicity." In *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CHI PLAY, article 235. https://doi.org/10.1145/3549498.

Shaw, A. (2012). "Do you identify as a gamer? Gender, race, sexuality, and gamer identity." *New Media & Society* 14, no. 1 (2012): 28–44. https://doi.org/10.1177/1461444811410394.

Sparrow, L., Butt, M.-A., Galwey, R., Jovic, D., and Hawwick, T. (2024). "Metaphors for moderating play: Rethinking the role of AI in online multiplayer games." Unpublished manuscript.

Xiao, S., Jhaver, S., & Salehi, N. (2022). "Addressing harm in online gaming communities: The opportunities and challenges for a restorative justice approach." *Journal of the ACM* 37(4), article 111. https://doi.org/10.1145/1122445.1122456.

Yang, Z., Grenan-Godbout, N., & Rabbany, R. (2023). Towards detecting contextual real-time toxicity for in-game chat. arXiv preprint arXiv:2310.18330.