Author/s:
Jovic, D;Galwey, R;Sparrow, LA;Butt, MA;Song, Y;Wang-Wills, S;Oliveira, EA

Title:
The AI Ally Project: Designing a Human-Centred AI Tool for Women's Safety in Online Spaces

Date:
2025-05-23

Citation:
Jovic, D., Galwey, R., Sparrow, L. A., Butt, M. A., Song, Y., Wang-Wills, S. & Oliveira, E. A. (2025). The AI Ally Project: Designing a Human-Centred AI Tool for Women's Safety in Online Spaces. Www Companion 2025 Companion Proceedings of the ACM Web Conference 2025, pp.2791-2795. ACM. https://doi.org/10.1145/3701716.3716878.

Persistent Link:
https://hdl.handle.net/11343/362036

The AI Ally Project: Designing a Human-Centred AI Tool for Women's Safety in Online Spaces

Dahlia Jovic
School of Culture and Communication
The University of Melbourne
Melbourne, Victoria, Australia
dahlia.jovic@unimelb.edu.au

Ren Galwey
School of Computing and Information Systems
The University of Melbourne
Melbourne, Victoria, Australia
ren.galwey@unimelb.edu.au

Lucy A. Sparrow
School of Computing and Information Systems
The University of Melbourne
Melbourne, Victoria, Australia
lucy.sparrow@unimelb.edu.au

Mahli-Ann Butt
School of Culture and Communication
The University of Melbourne
Melbourne, Victoria, Australia
mahliann.butt@unimelb.edu.au

Yige Song
School of Computing and Information Systems
The University of Melbourne
Melbourne, Victoria, Australia
yige.song1@unimelb.edu.au

Sable Wang-Wills
School of Computing and Information Systems
The University of Melbourne
Melbourne, Victoria, Australia
sable.w@unimelb.edu.au

Eduardo A. Oliveira
School of Computing and Information Systems
The University of Melbourne
Melbourne, Victoria, Australia
eduardo.oliveira@unimelb.edu.au

ABSTRACT

Online gender-based harassment is a widespread and persistent issue that disproportionately affects girls, young women, and gender diverse individuals. Systemic barriers, such as ineffective reporting mechanisms and fear of retaliation, often prevent victims and bystanders from taking action to address harm. These barriers emphasise the need for innovative, human-centred approaches to online safety. The AI Ally project is developing a trauma-informed, co-designed artificial intelligence (AI) tool that supports victims and witnesses of online harassment. The tool provides victims and witnesses with practical ways to document, reflect on, and report harassment, with the goal of empowering users and promoting safer digital environments.

This paper presents preliminary findings from the AI Ally project, drawing on over 230 survey responses from girls, young women, and gender diverse individuals living in Australia. We outline how survey insights, guided by a Feminist Participatory Action Research (FPAR) framework, are shaping the tool's design through a collaborative approach with participants. Additionally, we discuss how AI Ally's core functionalities and design considerations aim to bridge critical gaps in existing moderation systems and harassment reporting tools. This research contributes to ongoing discussions on ethical AI and technologies that are designed to be inclusive and responsive to diverse user needs.

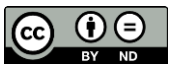
CCS CONCEPTS

•Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI

KEYWORDS

Online harassment; gender-based harassment; AI moderation; feminist participatory action research; human-centred design

ACM Reference format:
Dahlia Jovic, Ren Galwey, Lucy A. Sparrow, Mahli-Ann Butt, Yige Song, Sable Wang-Wills, Eduardo A. Oliveira. 2025. The AI Ally Project: Designing a Human-Centred AI Tool for Women's Safety in Online Spaces. In *Companion Proceedings of the ACM Web Conference 2025, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 5 pages.* https://doi.org/10.1145/3701716



WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04.
https://doi.org/10.1145/3701716

1 Introduction

1.1 Online Harassment

Plan International's *Free to Be Online* survey, which included over 14,000 girls in 32 countries, found that more than half of the respondents experienced online harassment [18]. In Australia, 65% of girls and young women reported experiencing online abuse—a higher rate than the global average [16]. While overall rates of digital harassment and abuse are similar for men and women, women are significantly more likely to face sexual harassment, reflecting gendered patterns of online abuse [3]. Similarly, sexually diverse and gender diverse individuals face elevated rates of online abuse that are compounded by intersecting forms of discrimination [2]. Online harassment manifests in various ways, including threatening language or images, persistent unwanted contact, location monitoring through social media, non-consensual sharing of private images, doxxing, and threats of violence [1]. The effects of gendered harassment are both deeply personal and far-reaching; many victims reduce their online engagement, self-censor, or withdraw from digital spaces entirely, limiting their ability to connect and express themselves freely in both public and private life [5].

The effects of online harassment underscore the need for comprehensive solutions that hold platforms and perpetrators accountable while supporting victims. In Australia, the Online Safety Act (2021) provides a regulatory framework designed to improve platform accountability while offering victims clearer pathways to report online harms [7]. The eSafety Commissioner can investigate complaints, issue takedown notices, and require platforms to implement measures like content removal and digital safety education [8]. Although regulatory initiatives provide an important foundation for combating online harassment, many victims still encounter challenges when navigating reporting pathways and seeking meaningful resolution [6].

As harmful content online continues to grow, social media and digital platforms are increasingly turning to automated and AI-driven moderation tools [9]. These tools are designed to filter harmful content on a large scale but may overlook context, intent, or evolving patterns of abuse. This can lead to user dissatisfaction and raises ethical concerns around privacy, autonomy, transparency, and fairness [10, 15, 17]. As emphasised in our research from the GAIM (Games and AI Moderation) project [14] and existing literature [21], punitive moderation systems that focus on flagging and penalising users often neglect the need for approaches that help individuals understand, challenge, and navigate moderation decisions.

1.2 Harassment Reporting Tools

Alongside the growing use of automation, several harassment reporting tools have incorporated features such as incident tracking and community-based peer assistance to respond to instances of abuse. Platforms like *Safecity* (Red Dot Foundation) [19] and *HarassMap* (Egypt) [13] provide individuals with a platform to anonymously report gender-based and sexual harassment. The collected data is used to map unsafe areas and identify patterns of abuse in physical locations. Similarly, *Harassment Manager* (Google) [11] was developed to assist journalists and public figures by consolidating reports of targeted abuse, organising flagged content, and offering tools for review. Right To Be's *Hate and Hope Tracker* [22] offers a peer-based network where volunteers assist victims of harassment through empathetic listening, advice, and strategies for managing their experiences. These platforms play a critical role in amplifying victims' voices and enabling broader interventions, yet gaps remain in equipping users with tools that allow them to make informed decisions about how to deal with harassment in digital settings [4]. This points to the need for co-designed online anti-harassment tools that reflect the lived experiences and needs of victims and marginalised groups.

1.3 The AI Ally Project

The AI Ally project—led by the University of Melbourne and funded by the eSafety Commissioner through the *Preventing Tech-based Abuse of Women Grants* Program—is co-designing a trauma-informed, AI-driven dashboard to support and empower girls, women and gender-diverse individuals aged 14-25 who have been victims and/or witnesses of online harassment. The dashboard aims to provide structured tools and processes for documenting, reflecting on, and reporting instances of abuse. This approach positions AI as an ally that moves away from punitive enforcement towards empowered action against harassment.

2 Methods

2.1 Survey Design

We conducted an online survey between May and August 2024, targeting girls, young women, and gender diverse individuals aged 14–25 in Australia who had experienced or witnessed online harassment. Ethics approval was obtained from the University of Melbourne, and trauma-informed training was provided to the research team by the Australian Childhood Foundation. We adopted a Feminist Participatory Action Research (FPAR) framework with the aim to amplify the voices of underrepresented and marginalised groups. The survey was distributed widely through social media platforms, advocacy groups, educational networks, and the project's Advisory Board, with a total 232 responses received.

Survey questions prompted participants to reflect on their encounters with online harassment, barriers to reporting, attitudes toward 'upstanding' (e.g., calling out harmful behaviour), and AI interventions. We asked participants optional questions about their responses to experiencing and witnessing harassment, focusing on common barriers to intervention. Additionally, the survey gathered diverse perspectives on how AI tools could better assist victims and encourage proactive actions

against harassment. To further inform AI Ally's design, we introduced participants to potential AI-driven functionalities (e.g., recording harassment, reporting it to moderators, messaging eSafety) and asked them to rank their perceived usefulness. Open-ended responses gave participants the space to share additional comments.

Participants who expressed interest in involvement beyond the survey were invited to subsequent evaluations and a 'Hackathon' (a collaborative event where users prototype and test new technology solutions) co-hosted with the project's industry partner, Girl Geek Academy. These sessions, planned for early 2025, will provide participants hands-on engagement with the AI Ally dashboard and give them the chance to explore its functions, assess its usability, and offer feedback that informs its refinement. These evaluations will assess how well the tool aligns with users' expectations and provide opportunities to expand its practical applications.

2.2 User Personas

We developed five user personas (Figure 1) based on aggregated survey data to better understand the target audience and inform the design of the AI Ally dashboard. These fictional profiles represented common characteristics, needs, and behavioural patterns among the target audience. Each persona provided an overview of online activities and interactions, a harassment-related incident prompting the use of the AI Ally dashboard, and a scenario demonstrating engagement with the tool's features. Key survey findings, including the emotional toll of harassment, limited confidence in reporting mechanisms, and a preference for supportive AI, guided decisions about the dashboard's functionality and accessibility. User personas continue to shape the iterative development of the tool, guiding refinements to its features and overall user experience.

Figure 1: AI Ally personas designed from survey responses. Ren Galwey/Research Rendered.

3 Survey Findings

Our survey revealed recurring patterns in how victims and witnesses experience and respond to online harassment, highlighting its profound impact. We found that online harassment was persistent and widespread, with 44% of respondents reporting they 'often' or 'always' encountered gendered harassment on at least one social media platform. Many respondents described harassment as a pervasive and routine part of their digital lives, often occurring across platforms and during everyday activities like posting photos or engaging in discussions. The most common forms included sexist comments (81%) and sexual harassment (63%). More complex manifestations of harassment like image-based abuse and AI-driven threats such as deepfakes were also reported, revealing the multifaceted and evolving nature of online abuse.

Although the reasons behind respondents' hesitations to report incidents of online harassment were not always clear, a few mentioned concerns about the effectiveness of platform responses. These included perceptions of inaction, inefficiency, and dismissiveness when complaints were filed. For example, one respondent noted that despite reporting extreme comments on TikTok, they did not believe the platform would block the offending accounts. Another respondent mentioned that Facebook did not consider a gender-based harassment event to have violated community guidelines. Actions involving external engagement, such as filing a report (46%) or discussing the incident with others (45%), were also less frequent compared to self-directed strategies like blocking or unfriending perpetrators (74%) and ignoring the harassment (56%)

Witnesses of online harassment (bystanders) often hesitated to intervene due to perceived risks, with 52% expressing concern for their personal safety and 47% fearing they might become targets themselves. Reasons for inaction included uncertainty about the context or people involved (45%) and not wanting to make the situation worse (44%). Respondents generally favoured indirect actions over confrontational approaches like messaging harassers or publicly calling out abuse. When introduced to a hypothetical AI Ally tool, the most preferred functions included documentation and structured reporting, such as generating harassment report files for submission to platform moderators and regulatory bodies like eSafety. These findings highlight the need for AI-driven interventions that provide practical, structured ways to manage harassment online while prioritising user safety and autonomy.

4 Discussion

4.1 The AI Ally Dashboard

Online gender-based harassment poses complex and pervasive challenges, with prior research revealing significant gaps in systems designed to address harm and empower both victims and bystanders. Studies indicate that inefficiency, lack of transparency, and limited feedback in online platform reporting systems erode user trust and deter individuals experiencing or witnessing harassment from acting [6, 14, 20, 21]. These barriers are intensified by fears of retaliation, judgment from others, and uncertainty about how to intervene [20]. While technical solutions like blocking and unfriending are often seen as emotionally helpful for users and provide a sense of control when formal mechanisms fall short [1, 12], these strategies are largely self-directed. In the survey findings, many respondents frequently

relied on these approaches rather than engaging with formal reporting systems. A smaller number of respondents raised concerns about the effectiveness of these systems, reflecting broader issues highlighted in the literature. Overall, the survey responses highlight how social, emotional, and systemic barriers often leave individuals to handle online harassment alone, emphasising the need for more supportive and trustworthy intervention mechanisms.

The AI Ally dashboard was developed in response to these challenges as a trauma-informed, co-designed AI tool that offers victims and witnesses structured ways to document, reflect on, and report harassment. The tool is being tested on Discord and uses Google's Perspective API to detect toxic messages, threats, and identity-based attacks, flagging harmful content for users to review. Rather than relying on automated enforcement, the tool enables users to decide how flagged content is handled. Key features include timestamped records, sender details, and message content, giving users the option to compile and download evidence-based reports. The dashboard provides a private space for users to track incidents and explore possible steps before deciding on a course of action. This approach aims to reduce the risks, emotional burdens, and uncertainty associated with managing online harassment while empowering users to make informed decisions about how to respond.

4.2 Limitations and Future Work

Developing the AI Ally dashboard has highlighted the complexities of creating an AI-driven tool that effectively responds to online harassment. Google's Perspective API provides key capabilities for detecting harmful language and assessing message toxicity. However, it faces challenges in identifying more subtle forms of harassment, such as sarcasm or indirect threats, as well as context-dependent abuse. Refinements to the AI model will involve integrating advanced techniques like conversational analysis and contextual language modelling, as well as collaborations with academic researchers, industry experts, and platform providers to improve detection capabilities. Real-world user testing will refine the dashboard to improve its usability, responsiveness, and effectiveness in helping users navigate online harassment. Ethical considerations, including transparency, user control over data, and fostering trust, will remain central to AI Ally's future iterations.

5 Conclusion

The AI Ally project calls for urgent and ethical solutions to online harassment, which disproportionately affects girls, young women, and gender diverse individuals. Grounded in the perspectives of victims and witnesses, the project addresses significant gaps in how harassment is managed online, including unclear reporting processes, lack of accountability from platforms, and the normalisation of harmful behaviours. Insights from the survey continue to inform and shape the design of the AI Ally dashboard, which offers structured tools for documentation, guided reporting, and user-controlled workflows. The project represents progress toward creating safer and more inclusive online spaces, with ongoing development shaped by participant contributions, collaboration with key stakeholders, and a focus on encouraging meaningful action in response to online harassment.

ACKNOWLEDGMENTS

This project was funded through the eSafety Commissioner's *Preventing Tech-based Abuse of Women Grants* Program – an Australian Government initiative. We extend our heartfelt thanks go to our industry partner, Girl Geek Academy, for their collaboration and support. We also thank our Advisory Board for their guidance and Maddy Weeks, our dedicated social media officer, for amplifying awareness about the project.

REFERENCES

[1] Amnesty International. 2018. Australia: Poll reveals alarming impact of online abuse against women. Retrieved February 2025 from https://www.amnesty.org.au/australia-poll-reveals-alarming-impact-online-abuse-women.

[2] Anastasia Powell, Adrian J. Scott, and Nicola Henry. 2020. Digital harassment and abuse: Experiences of sexuality and gender minority adults. *European Journal of Criminology* 17, 2, 199–223. https://doi.org/10.1177/1477370818788006.

[3] Anastasia Powell and Nicola Henry. 2015. Digital harassment and abuse of adult Australians: A summary report. Retrieved February 2025 from https://www.parliament.nsw.gov.au/lcdocs/other/7351/Tabled%20Document%20-Digital%20Harassment%20and%20Abuse%20of%20A.pdf.

[4] Anna Wojtowicz, Graham J. Buckley, and Sandro Galea. 2024. Chapter 7: Online harassment. In *Social Media and Adolescent Health* (A. Wojtowicz, G. J. Buckley, and S. Galea, Eds.). National Academies Press, Washington, DC.

[5] Bridget Harris and Laura Vitis. 2020. Digital intrusions: Technology, spatiality, and violence against women. *Journal of Gender-Based Violence* 4, 3, 325–341. https://doi.org/10.1332/239868020X15986402363663.

[6] Chandell E. Gosse. 2021. More barriers than solutions: Women's experiences of support with online abuse. Ph.D. Dissertation. University of Western Ontario, London, Canada. *Electronic Thesis and Dissertation Repository*, 7628. Retrieved February 2025 from https://ir.lib.uwo.ca/etd/7628.

[7] eSafety Commissioner. Learn about the Online Safety Act. Retrieved February 2025 from https://www.esafety.gov.au/newsroom/whats-on/online-safety-act.

[8] eSafety Commissioner. Regulatory Guidance. Retrieved February 2025 from https://www.esafety.gov.au/industry/regulatory-guidance.

[9] Emma Llansó, Joris van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. Artificial Intelligence, Content Moderation, and Freedom of Expression. TWG: Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Netherlands. Retrieved February 2025 from https://coilink.org/20.500.12592/37pvprf. COI: 20.500.12592/37pvprf.

[10] European Union Agency for Fundamental Rights. 2023. Online content moderation: Current Challenges in Detecting Hate Speech. Publications Office of the European Union, Luxembourg. Retrieved February 2025 from https://fra.europa.eu/sites/default/files/fra_uploads/fra-2023-online-content-moderation_en.pdf.

[11] Google. Harassment Manager. Retrieved February 2025 from https://jigsaw.google.com/harassment-manager.

[12] Hanka Machackova, Alice Cerna, Andrea Sevcikova, Lucie Dedkova, and Kjell Daneback. 2013. Effectiveness of coping strategies for victims of cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 7, 3, Article 5. https://doi.org/10.5817/CP2013-3-5.

[13] HarassMap. HarassMap. Retrieved February 2025 from https://harassmap.org/en.

[14] Lucy A. Sparrow, Ren Galwey, Dahlia Jovic, Taylor Hardwick, and Mahli-Ann Butt. 2024. Towards ethical AI moderation in multiplayer games. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY, Article 344 (October 2024), 32 pages. https://doi.org/10.1145/3677109.

[15] María D. Molina and S. Shyam Sundar. 2022. When AI moderates online content: Effects of human collaboration and interactive transparency on user

trust. *Journal of Computer-Mediated Communication* 27, 4 (July 2022), Article zmac010. https://doi.org/10.1093/jcmc/zmac010.

[16] Norman Hermant. 2020. Young Australian women cop more online harassment than global average, report finds. ABC News. Retrieved February 2025 from https://www.abc.net.au/news/2020-10-05/young-australian-women-online-abuse-harassment-planinternational/12725286.

[17] Oscar Gladwin. 2024. Navigating ethical dilemmas in AI content moderation: Balancing privacy, free speech, and fairness. Retrieved February 2025 from https://doi.org/10.13140/RG.2.2.18312.23042.

[18] Plan International. 2020. Free to be online: Girls and young women's experiences of online harassment. Plan International, Surrey. Retrieved February 2025 from https://www.plan.org.au/publications/free-to-be-online.

[19] Red Dot Foundation. Safecity. Retrieved February 2025 from https://webapp.safecity.in.

[20] Randy Yee Man Wong, Christy Cheung, Bo Xiao, and Jason Thatcher. 2020. Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. *Information Systems Research* 32, 1, Article 5. https://doi.org/10.1287/isre.2020.0983.

[21] Renkai Ma, 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, CSCW2 (2023), 1–30. https://doi.org/10.1145/3610069.

[22] Right To Be. Hate and Hope Tracker. Retrieved February 2025 from https://hateandhope.righttobe.org/pages/about-page.