Envisioning Responsible Personal Moderation for Gender-Based Online Harm

ANNA RICARDA LUTHER, Institute for Information Management Bremen GmbH, Germany and University of Bremen, Germany

Gender-based hate speech remains pervasive on social media platforms. While platforms primarily rely on highly automated, platform-wide moderation systems, these approaches often fail to address the contextual and subjective nature of gender-based harms. In this position paper, I argue that personal moderation tools can play a crucial complementary role in mitigating these harms. Drawing on findings from a Delphi study with users targeted by hate speech, I highlight a misalignment between current personal moderation tools and the needs of those they are intended to support. Effective personal moderation should be designed by centering affected users' perspectives and must empower users without shifting responsibility for safety away from platforms.

Additional Key Words and Phrases: Hate Speech Interventions, Social Media, Personal Moderation, Activism

ACM Reference Format:
Anna Ricarda Luther. 2026. Envisioning Responsible Personal Moderation for Gender-Based Online Harm. 1, 1 (March 2026), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Gender-Based Harms Online and the Role of Personal Moderation

Hate speech constitutes a massive issue on social media platforms. According to a representative survey, 50% of young adults in Europe have encountered hate speech on social media [5]. Hate speech disproportionately affects marginalized groups, e.g. those marginalized based on gender [12–14]. Exposure to hate speech has been shown to have a silencing effect by deterring targets from participating in public discourse or disclosing parts of their identities. For example, in Europe, 52% of women report that they voice their opinions less on social media due to fear of encountering hate speech [5]. But the consequences of hate speech extend beyond digital spaces, posing real-world risks to mental and physical well-being [17]. Crucially, hate speech also correlates with heightened risks to physical safety, as it is deeply intertwined with real-world violence against targeted communities [2, 19]. Our own work, in which we interviewed queer-feminist activists (among others), paints a similar picture, as one participant described *"I also see a danger that it [the negative comments they get during activist actions] could become physical"* [10]. This motivated us to develop new content moderation solutions to more effectively combat hate speech.

To address hate speech online, most platforms deploy platform-wide content moderation systems that aim to remove violating content or sanction users who repeatedly breach platform guidelines [7, 11, 18]. These systems rely heavily on automated hate speech detection models, which have been shown to exhibit systematic biases [1, 3]. While automated systems are becoming increasingly effective at detecting explicit hate speech directed at women, they perform substantially worse when addressing hate speech targeting LGBTQIA+ communities and implicit or context-dependent forms of hate speech across groups [4]. Gender-based harassment often relies on coded language,

Author's Contact Information: Anna Ricarda Luther, aluther@ifib.de, Institute for Information Management Bremen GmbH, Bremen, Germany and University of Bremen, Bremen, Germany.

Manuscript submitted to ACM

contextual cues, or cumulative microaggressions that constitute implicit hate speech [15, 16]. Thus, automated hate speech detection systematically fails to capture the subjective and contextual nature of gender-based harms online. Personal moderation tools are user-controlled features that allow individuals to shape the content they encounter by setting account-specific preferences [9], thereby introducing the contextual input necessary to address gender-based harms. In this way, personal moderation could serve as a complementary layer of protection that enables users to proactively shape safer online environments and sustain participation in public discourse.

2 Envisioning Personal Moderation Beyond Platform Logics

For personal moderation to effectively mitigate gender-based harms online, these tools must be designed in close collaboration with those affected by such harms. Our broader project initially set out to explore the design of alternative social media spaces; however, early interviews with activists quickly revealed that gender-based harassment was not a peripheral issue but a central barrier to participation. Activists described how existing moderation systems routinely failed to address the contextual and cumulative nature of abuse, motivating us to more systematically investigate what meaningful personal moderation would require [10]. To this end, we conducted a Delphi study bringing together social media users targeted by hate speech. Through iterative rounds of structured reflection and consensus-building, our study surfaced shared priorities and design considerations for personal moderation tools that are largely absent from current platform implementations [8]. Participants consistently emphasized that harm is highly contextual and unevenly distributed across different parts of a platform, challenging one-size-fits-all moderation approaches. Our findings reveal a clear mismatch between the needs articulated by users experiencing hate speech and the personal moderation options currently available on major platforms. These insights demonstrate that effective personal moderation depends on centering the perspectives of those the tools are meant to serve.

3 Personal Moderation vs. Platform Responsibility

While personal moderation tools can reduce exposure to gender-based harms, they must not shift the responsibility for safety from platforms to users. Framing personal moderation as an individual responsibility risks further burdening those already disproportionately affected by online abuse and may normalize the persistence of harmful environments [6]. Personal moderation should instead be understood as an empowering complement to platform-level moderation, not a substitute for it. Platforms remain responsible for preventing, detecting, and responding to abuse at scale; personal moderation should serve to enhance users' autonomy and resilience within these systems. When designed accordingly, personal moderation can function as one building block in addressing gender-based online harm without reinforcing the structural inequalities that give rise to gender-based harms in the first place.

4 Conclusion

Gender-based harms online are pervasive and deeply contextual, yet dominant moderation approaches continue to privilege platform-wide, automated solutions that insufficiently address these realities. In this position paper, we argue that personal moderation tools can serve as a critical complementary layer in mitigating gender-based harm, given they are designed by centering the perspectives of those most affected and implemented without shifting responsibility from platforms to users. Personal moderation should empower users to navigate online spaces more safely while remaining embedded within broader systems of platform accountability. Future research and design efforts must therefore treat personal moderation not as an individualized fix to structural problems, but as one component of a multi-layered approach to addressing gender-based harms online.

References

[1] Melisa Castellanos, Alexander Wettstein, Sebastian Wachs, Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, and Ludwig Bilz. 2023. Hate speech in adolescents: A binational study on prevalence and demographic differences. In *Frontiers in Education*, Vol. 8. Frontiers Media SA, Lausanne, Switzerland, 1076249.

[2] Esli Chan. 2023. Technology-Facilitated Gender-Based Violence, Hate Speech, and Terrorism: A Risk Assessment on the Rise of the Incel Rebellion in Canada. *Violence Against Women* 29, 9 (2023), 1687–1718. arXiv:https://doi.org/10.1177/10778012221125495 doi:10.1177/10778012221125495 PMID: 36226437.

[3] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 12 pages. doi:10.1145/3689904.3694704

[4] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in moderation: How commercial content moderation apis over-and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–26.

[5] HateAid and The Landecker Digital Justice Movement. 2021. Grenzenloser Hass im Internet – Dramatische Lage in ganz Europa. https://hateaid.org/wp-content/uploads/2022/04/HateAid-Report-2021-DE.pdf

[6] Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2025. "Ignorance is not Bliss": Designing Personalized Moderation to Address Ableist Hate on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 173, 18 pages. doi:10.1145/3706598.3713997

[7] Reddit Inc. 2025. *Reddit Rules (Content Policy)*. Reddit Inc. https://redditinc.com/policies/content-policy Retrieved: 2025-11-20.

[8] Instagram Help Center. 2025. Privacy Settings & Information on Instagram. Available at: https://help.instagram.com/196883487377501/. Retrieved: 2025-09-04.

[9] Shagun Jhaver and Amy X Zhang. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 27, 5 (2023), 14614448231217993.

[10] Anna Ricarda Luther, Hendrik Heuer, Stephanie Geise, Sebastian Haunss, and Andreas Breiter. 2025. Social Media for Activists: Reimagining Safety, Content Presentation, and Workflows. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 956, 18 pages. doi:10.1145/3706598.3713351

[11] Meta. 2025. *Restricting Accounts – Enforcement Transparency*. Meta. https://transparency.meta.com/enforcement/taking-action/restricting-accounts/ Retrieved: 2025-11-20.

[12] Sandra Miranda, Fabio Malini, Branco Di Fátima, and Jorge Cruz. 2022. I love to hate!: the racist hate speech in social media. *European Conference on Social Media* (2022). doi:10.34190/ecsm.9.1.311

[13] Sandra Lopes Miranda. 2023. Analyzing Hate Speech Against Women on Instagram. *Open Information Science* 7 (2023). doi:10.1515/opis-2022-0161

[14] Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Inf. Process. Manag.* 57 (2020), 102087. doi:10.1016/J.IPM.2019.102087

[15] Arianna Muti. 2025. *Hidden in plain sight: detecting misogyny beneath ambiguities and implicit bias in language*. Ph. D. Dissertation. alma. https://amsdottorato.unibo.it/id/eprint/12195/

[16] Louise Richardson-Self. 2018. Woman-Hating: On Misogyny, Sexism, and Hate Speech. *Hypatia* 33, 2 (2018), 256–272. doi:10.1111/hypa.12398

[17] Francesca Stevens, Jason RC Nurse, and Budi Arief. 2021. Cyber stalking, cyber harassment, and adult mental health: A systematic review. *Cyberpsychology, Behavior, and Social Networking* 24, 6 (2021), 367–376.

[18] TikTok Safety Center. 2025. *Content Violations & Bans – Account and User Safety*. TikTok. https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans Retrieved: 2025-11-20.

[19] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-Black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology* 60, 1 (2020), 93–117.