New Opportunities, Risks, and Harm of Generative AI for Fostering Safe Online Communities

Guo Freeman
guof@clemson.edu
Clemson University
USA

Douglas Zytko
dzytko@umich.edu
University of Michigan-Flint
USA

Afsaneh Razi
afsaneh.razi@drexel.edu
Drexel University
Philadelphia, PA, USA

Cliff Lampe
cacl@umich.edu
University of Michigan
Ann Arbor, MI, USA

Heloisa Candello
heloisacandello@br.ibm.com
IBM Research
São Paulo, SP, Brazil

Timo Jakobi
timo.jakobi@th-nuernberg.de
Nuremberg University of Applied Sciences Georg Simon Ohm
Nuremberg, Germany

Konstantin "Kosta" Aal
konstantin.aal@uni-siegen.de
University of Siegen
Siegen, Germany

ABSTRACT

Recently, there is a growing trend of using generative AI systems and tools for fostering and protecting online collaborative communities. Yet, existing AI tools may introduce new risks and even harm to diverse communities' online safety. How to better maximize the novel opportunities of AI and mitigate its emerging risks and harm for our future online safety is a critically needed discussion for the HCI community. Featuring experts from both industry and academia, the goal for this panel is to promote interdisciplinary, community-wide discussions and collective reflections on important questions and considerations at the unique intersection of AI and online communities, including but not limited to: how the design of AI systems may discourage existing online harm but also invite new online harm in various online spaces; how different populations, cultures, and communities may perceive and experience AI's new roles for their online safety; and what new strategies, principles, and directions can be envisioned and identified to better design future AI technologies to protect rather than harm various online communities.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Human computer interaction (HCI)**.

KEYWORDS

generative AI, online communities, online safety

GROUP Companion '25, January 12–15, 2025, Hilton Head, SC, USA

ACM ISBN 979-8-4007-1187-9/25/01.
https://doi.org/10.1145/3688828.3700747

ACM Reference Format:
Guo Freeman, Douglas Zytko, Afsaneh Razi, Cliff Lampe, Heloisa Candello, Timo Jakobi, and Konstantin "Kosta" Aal. 2025. New Opportunities, Risks, and Harm of Generative AI for Fostering Safe Online Communities. In *The 2025 ACM International Conference on Supporting Group Work (GROUP Companion '25), January 12–15, 2025, Hilton Head, SC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3688828.3700747

1 INTRODUCTION

Artificial Intelligence (AI) has become the core driver of innovation in emerging technologies [18] and is increasingly being used to support our everyday lives across a variety of sectors. More recently, there is a growing trend of using generative AI (i.e., AI systems that are often trained using large datasets and can generate new content based on learned and reproduced patterns from the training datasets [5]) systems and tools for creating safe online environments. Despite these new opportunities, some other research has shown that existing AI tools may introduce new risks and even harm to diverse communities' online safety. In the following, we will outline some specific examples of these new opportunities, risks, and harm of leveraging AI for online safety.

AI for moderating online harassment and cyberbullying. A typical example of using AI for online safety is to use AI-based moderation to monitor, detect, and mitigate online harassment and cyberbullying. For instance, AI has been used to automatically filter certain keywords to block posts or comments that include specific harassing terms and phrases, such as the AutoModerator bot on Reddit [4, 7, 9, 14, 17, 22] and flagging systems used in gaming [16, 26]. Machine learning-trained AI can also detect toxicity in game communication [21, 25] or conduct voice analysis to detect sexual harassment online by searching for clues of fear, anger, and disgust emotions in women's voices [15, 24, 26]. However, AI-based online content moderation has also been criticized for disproportionately targeting marginalized individuals (e.g., women, people with mental illness, and Black individuals), such as further marginalizing those with eating disorders by reasserting certain bodies as not

valuable and thus worthy of censorship [10]; unjustifiably removing content created by transgender and Black users expressing their marginalized identities [12]; and censoring women and vulnerable users while uplifting the voices of online abusers [3].

AI for detecting and mitigating sexual risks. Beyond the traditional social media context, AI has been used for detecting and mitigating sexual risks in online dating [8, 30]. For example, Bumble has released an open-source "cyberflashing" detection AI [28]. Examples beyond online dating include skin detection AI to identify non-consensual and illegal sexual imagery [23, 27] as well as detection of child grooming and sex trafficking through social media [2, 29]. Despite this growing interest in AI for preventing sexual harm, recent literature shows that sexual risk detection AI models are commonly trained on datasets of publicly available social media data (e.g., comments on public posts) or police officers impersonating children to catch child predators [20]. There persists an absence of ecologically valid data around private messaging interactions and physical sexual experiences following from online discovery, in which recently researchers started to address such as identification of and understanding the language around sexual risks within youths' private conversations [19].

AI for youth online safety. As mentioned above, adolescents, a particularly vulnerable population of technology users, face various online risks including sexual harm. Therefore, there have been various research efforts to utilize AI to detect and mitigate online risks for adolescents [20]. For instance, researchers have utilized multimodal approaches to detect unsafe interactions [1] and found that meta-data from youth's private conversations provides sufficient evidence for machine learning models to detect unsafe interactions. This means that with minimum information, privacy-invasive tools such as parental monitoring apps [11] or prioritizing security over children's safety by implementing end-to-end encryption [1] could be addressed. However, with the recent introduction of AI-based conversational agents (CAs) such as chatGPT, adolescents have sought knowledge and support through these chatbots on sensitive topics, leading to potential AI-introduced online risks. For example, it's been documented that these chatbots expose teens to sexually inappropriate content, false information, or misleading and potentially harmful advice [2].

AI that causes new online harm. Additionally, AI technologies could be used to cause new online harm rather than mitigating such harm. One example of this is using AI intentionally for is real-time human-bot coordinated group attacks in live streaming communities (e.g, "hate raids," see [6, 13]). In this case, AI is intentionally used to perform new online attacks at a larger scale and at a much faster pace that goes beyond the capacity of existing traditional harm mitigation approaches (e.g., human-based moderation). For instance, massive bot accounts start to follow and/or unfollow a streamer to create the notification sound to disrupt people's streaming and viewing experience. These bots can also be used to overwhelm the live chat by generating a large amount of hate messages within a very short time frame that a human moderator is unable to manage [6, 13].

[1]https://www.itpro.com/security/encryption/359943/what-is-end-to-end-encryption-and-why-is-everyone-fighting-over-it
[2]https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/

Therefore, we believe that how to better maximize the novel opportunities of AI and mitigate its emerging risks and harm for our future online communities is a critically needed discussion for the HCI community. Featuring experts from both industry and academia, the goal for this panel is to promote interdisciplinary, community-wide discussions and collective reflections.

2 PANEL THEMES AND GOALS

In this panel, we aim to discuss important questions and considerations at the unique intersection of generative AI, online communities, and cybersecurity, including but not limited to: how the design of AI systems may discourage existing online harm but also invite new online harm in various online spaces; how different populations, cultures, and communities may perceive and experience AI's new roles for their online safety; and what new strategies, principles, and directions can be envisioned and identified to better design future AI technologies to protect rather than harm various online communities. Some example topics and questions we will focus on in this panel include:

- What are some existing methods and practices of using AI to protect diverse communities' online safety and what are the limitations/challenges of these methods and practices?
- What are some forms of new online risks and harm that AI may cause for diverse online communities rather than protecting them?
- How, if at all, do these new risks and harm for people's online safety caused by AI also be translated to offline risks and harm?
- Why do some people consider AI as negative or harmful for their online safety while others do not?
- How can we better understand and approach AI's unique challenges for online safety across various sociocultural contexts (e.g., in the global south context and across different age groups)?
- How can we identify new approaches, designs, and directions to create future AI technologies to maximize their opportunities while mitigating their risks and harm for people's online safety?

The panel format will alternate between the moderator posing questions to the panelists, taking questions from the audience, and posing questions to the audience, to gather community opinions.

3 PANELISTS AND MODERATOR

Panelists have expertise in the area of AI for online safety across various online contexts (e.g., games, live streaming, XR, social media, and online dating) with diverse research approaches (e.g., qualitative, experimental, design, and machine learning). Taken together, panelists bring unique knowledge and vision from both academia and industry to this topic.

Guo Freeman (panelist) is an Associate Professor of Human-Centered Computing at Clemson University. Her work focuses on how interactive technologies such as digital games, live streaming, social VR, and AI shape interpersonal relationships and group behavior; and how to design safe, inclusive, and supportive social VR spaces to mitigate emergent harassment risks, such as through AI-based moderation.

Douglas Zytko (panelist) is an Associate Professor and Director of Graduate Studies in the College of Innovation & Technology at the University of Michigan-Flint. His research uses consent as a lens to understand and design mitigative solutions for sexual harm. His work explores data donation of online dating sexual experiences as a human-centered approach to improving sexual risk detection AI.

Afsaneh Razi (panelist) is an Assistant Professor at Drexel University's Department of Information Science. Her research area is positioned at the intersection of HCI and AI to address sociotechnical issues. Specifically, her work aims to address the critical and timely problem of online safety by leveraging a multidisciplinary approach of human-centered AI to characterize and identify risks vulnerable users encounter online and develop online safety interventions.

Cliff Lampe (panelist) is a Professor and Associate Dean for Academic Affairs at the University of Michigan School of Information. He also serves as Chair of the CHI Steering Committee and is a member of the CHI Academy. His work has highlighted moderation in online spaces, the effects of harassment and disinformation in online spaces, and recently the motivations behind harassers in these spaces.

Heloisa Candello (panelist) is a research scientist at the Responsible &Inclusive group at IBM Research laboratory. She has experience conducting mixed-methods research in the collection, design, and evaluation of conversational systems. Her research resulted in several publications in leading conferences (CHI, CSCW, DRS, DUXU) and recognition in the HCI and Design field.

Timo Jakobi (panelist) is a professor whose work focuses on making abstract protection concepts accessible to companies, demonstrating how data protection and responsible AI can serve customer interests and provide added value. His research bridges legal frameworks with human-computer interaction methods to make digital legislation more empirically grounded, aiming to reduce compliance uncertainty and help businesses strategically use data protection as a value proposition in customer communications.

Konstantin Aal (moderator) is a PostDoc at the Chair of Information Systems and New Media at the University of Siegen. His research focuses on the use of social media by political activists in conflict areas such as Palestine, Iran, Tunisia and Syria. His recent publications revolve around the idea of the personalised AI companion and how it can be used to augment rather than replace the user.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under award 2112878 and 2342393.

REFERENCES

[1] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–30.

[2] Philip Anderson, Zheming Zuo, Longzhi Yang, and Yanpeng Qu. 2019. An Intelligent Online Grooming Detection System Using AI Technologies. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858973

[3] Carolina Are. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies* 20, 5 (2020), 741–744.

[4] Hannah Bloch-Wehba. 2020. Automation in moderation. *Cornell Int'l LJ* 53 (2020), 41.

[5] Alice Cai, Steven R Rick, Jennifer L Heyman, Yanxia Zhang, Alexandre Filipowicz, Matthew Hong, Matt Klenk, and Thomas Malone. 2023. DesignAID: Using Generative AI and Semantic Diversity for Design Inspiration. In *Proceedings of The ACM Collective Intelligence Conference*. 1–11.

[6] Jie Cai, Sagnik Chowdhury, Hongyang Zhou, and Donghee Yvette Wohn. 2023. Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities. *arXiv preprint arXiv:2305.16248* (2023).

[7] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*. 121–126.

[8] Isha Datey, Hanan Khalid Aljasim, and Douglas Zytko. 2022. Repurposing AI in Dating Apps to Augment Women's Strategies for Assessing Risk of Harm. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, Taiwan) *(CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 150–154. https://doi.org/10.1145/3500868.3559472

[9] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 11–17.

[10] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[11] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. 2018. Safety vs. surveillance: what children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[12] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[13] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T Hancock, and Zakir Durumeric. 2023. Hate raids on Twitch: Echoes of the past, new modalities, and implications for platform governance. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.

[14] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[15] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[16] Yubo Kou 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.

[17] Emma Llansó, Joris Van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. Artificial intelligence, content moderation, and freedom of expression. (2020).

[18] Yang Lu. 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics* 6, 1 (2019), 1–29.

[19] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2023. Sliding into my DMs: Detecting uncomfortable or unsafe sexual risk experiences within Instagram direct messages grounded in the perspective of youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[20] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (oct 2021), 38 pages. https://doi.org/10.1145/3479609

[21] Elizabeth Reid, Regan L Mandryk, Nicole A Beres, Madison Klarkowski, and Julian Frommel. 2022. "Bad vibrations": Sensing toxicity from in-game audio features. *IEEE Transactions on Games* 14, 4 (2022), 558–568.

[22] Kim Renfro. 2016. For whom the troll trolls: A day in the life of a Reddit moderator. *Business Insider* (2016).

[23] Napa Sae-Bae, Xiaoxi Sun, Husrev T. Sencar, and Nasir D. Memon. 2014. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*. 5332–5336. https://doi.org/10.1109/ICIP.2014.7026079

[24] Shikhar Sakhuja and Robin Cohen. 2020. RideSafe: Detecting Sexual Harassment in Rideshares. In *Canadian Conference on Artificial Intelligence*. Springer, 464–469.

[25] Kelsea Schulenberg, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, and Nathan J. McNeese. 2023. Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. In *Proceedings of the 2023 CHI Confrence on Human Factors in Computing Systems*.

[26] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*. 19–24.

[27] Muhammad Uzair Tariq, Afsaneh Razi, Karla A. Badillo-Urquiola, and Pamela J. Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In *Interacción*.

[28] Rachel Thompson. 2022. Bumble makes cyberflashing detection tool available as open-source code. *Mashable* (Oct 2022). https://mashable.com/article/bumble-cyberflashing-private-detector-open-source

[29] Hao Wang, Congxing Cai, Andrew Philpot, Mark Latonero, Eduard H. Hovy, and Donald Metzler. 2012. Data Integration from Open Internet Sources to Combat Sex Trafficking of Minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (College Park, Maryland, USA) *(dg.o '12)*. Association for Computing Machinery, New York, NY, USA, 246–252. https://doi.org/10.1145/2307729.2307769

[30] Wenqi Zheng, Emma Walquist, Isha Datey, Xiangyu Zhou, Kelly Berishaj, Melissa Mcdonald, Michele Parkhill, Dongxiao Zhu, and Douglas Zytko. 2024. "It's Not What We Were Trying to Get At, but I Think Maybe It Should Be": Learning How to Do Trauma-Informed Design With a Data Donation Platform for Online Dating Sexual Violence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642045