Novel Approaches for Understanding and Mitigating Emerging New Harms in Immersive and Embodied Virtual Spaces: A Workshop at CHI 2024

Guo Freeman
guof@clemson.edu
Clemson University
USA

Julian Frommel
j.frommel@uu.nl
Utrecht University
Netherlands

Regan L. Mandryk
reganmandryk@uvic.ca
University of Victoria
Canada

Jan Gugenheimer
jan.gugenheimer@tu-darmstadt.de
TU-Darmstadt/Telecom Paris
France

Lingyuan Li
lingyu2@g.clemson.edu
Clemson University
USA

Daniel Johnson
dm.johnson@qut.edu.au
Queensland University of Technology
Australia

ABSTRACT

As online spaces facilitate increasingly immersive and embodied experiences, concerns about how these emerging spaces may amplify and extend existing online harms and even lead to new harms, and how HCI researchers and developers can work to mitigate such harms also grow. Typical examples of these new and understudied forms of harm range from embodied harassment in social Virtual Reality (VR) to racist Zoombombing, new AI-powered online attacks such as hate raids on Twitch, and harmful virtual world design to manipulate users. This workshop aims to bring together a set of interdisciplinary researchers and practitioners from HCI and adjacent fields to explore further how these new harms continue to shape the current research discourse of online safety, cybersecurity, and immersive and embodied interactions in HCI, and to collectively identify what new technologies and mechanisms can be envisioned, designed, and implemented to better understand and mitigate these harms.

CCS CONCEPTS

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

KEYWORDS

toxicity, online harm, online safety, harassment, embodiment, harm mitigation, artificial intelligence, immersive virtual worlds

ACM Reference Format:
Guo Freeman, Julian Frommel, Regan L. Mandryk, Jan Gugenheimer, Lingyuan Li, and Daniel Johnson. 2024. Novel Approaches for Understanding and Mitigating Emerging New Harms in Immersive and Embodied Virtual Spaces: A Workshop at CHI 2024. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3636288

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3636288

1 BACKGROUND AND MOTIVATION

Trigger Warning: In this workshop description, we explain emerging new harms in various immersive and embodied virtual spaces, such as: simulated physical violence, embodied harassment, zoombombing, hate raids on Twitch, and harmful user- or developer-generated virtual spaces/applications.

How diverse online users encounter, experience, and manage various forms of toxic, problematic, and harmful interactions (e.g., harassment, verbal abuse, hate speech, and flaming on social media platforms and online forums) remains a severe and pervasive issue in today's online social spaces, which seriously damages victims' mental, emotional, and physical well-being [6, 22, 51]. Moving towards gaming and virtual world contexts paints an even more direct picture. Indeed, as early as in 1993, Dibbell's "A Rape In Cyberspace" detailed sexual assaults of women users in one of the first virtual worlds called LambdaMOO [13]. Now toxicity is largely accepted and normalized in online gaming [5, 21, 54] with statistics suggesting that 83% of adult gamers experience harassment in online multiplayer games with severe effects like causing distress [3]. With the rise of new immersive technologies such as Augmented and Virtual Reality which heavily rely on immersion, presence and embodiment, these toxic behaviours have the potential to be amplified or even take completely new shapes. Typical examples of these new and understudied forms of harm in immersive and embodied virtual spaces range from *embodied harassment* in social Virtual Reality (VR) [20] to new forms of online racism such as *racist Zoombombing* [34] and *Ugandan Knuckles* in VRChat [2], new AI-powered online attacks such as *hate raids* on Twitch [11], user- or developer-generated harmful design to manipulate people's actions [29], and new types of perceptual manipulations leveraging the core abilities of XR technology to create illusions to the user [10, 24, 50]. In the following, we will outline some specific examples of these new potential harms.

New Embodied Harassment. In social VR, multiple users can interact with one another through VR head-mounted displays in 3D virtual spaces while also leveraging other technological features (e.g., partial-to-fully body-tracked avatars, predominate voice communication, and body language and gestures) to simulate offline-like social interactions (e.g., touching and grabbing others) [19, 32].

These unique features thus allow users to meet, interact, and socialize in more embodied (i.e., experiencing a virtual body representation as one's own [45]) and immersive (i.e., being enveloped by, included in, and interacting with the virtual environment [55]) ways than in other, screen-mediated online social spaces (e.g., social media and gaming). However, this focus on embodied and immersive experiences has also led to intensified and more physicalized forms of harassment in social Virtual Reality (VR) compared to other online contexts, ranging from trash talking women, drawing penises, and virtual "groping" to the most recent "rape" in the metaverse [7, 8, 16, 17, 20, 35, 39–43, 46, 47, 56]. Prior work in HCI has thus identified several new characteristics of online harassment in social VR and highlighted this type of *embodied harassment* as an emerging but understudied form of harassment in novel online social spaces [20]. In this context, harassing behaviors are both conducted and experienced through a sense of embodiment about one's virtual body with a higher awareness of body ownership and more physical and transformative interactive experiences [45].

New Forms of Online Racism. As HCI continues to advocate an anti-racism/racist research agenda [1, 15], significant works have begun to explain how social technologies may enable various new forms of racism against non-white populations in ways that are "inherent and foundational to Internet and gaming cultures" [34]. These explorations have uncovered several insidious patterns of racial abuse, including how African American gamers' bodies are labeled as deviant in online gaming (e.g., Xbox Live) through a process of questioning, provoking, instigating, and ultimately racism [23], occurrences of online harassment of Asian Americans during COVID-19 [49], and incidents of racist Zoombombing (i.e., using Zoom, the videoconferencing software, to attack unsuspecting users with racist content [34]). As racism remains a severe and pervasive issue in today's world, these works thus highlight the urgent need for designing new socio-technical platforms for racial minorities in a white-dominated society to better cope with interpersonal racism-based harm both online and offline [48].

New AI-powered Online Attacks. Advanced Artificial Intelligence (AI) technologies, such as AI-based moderation [26, 33, 52, 53], are often considered an effective approach to help mitigate online harms by automatically filtering certain keywords to block posts or comments that include specific harassing terms and phrases (e.g., the AutoModerator bot on Reddit [9, 12, 14, 27, 30, 37]) or by using Natural Language Processing techniques to automatically detect cyberbullying content [4, 38]. Yet, if not used appropriately with caution, AI technologies could also be used to cause new online harms rather than mitigating such harms. One most recent example is the so-called "hate raids" in live streaming communities, which is a form of human-bot coordinated group attack in real-time [11, 25]. During such a "hate raid," massive bot accounts start to follow and/or unfollow a given streamer to intentionally create the notification sound, which significantly harms people's streaming and viewing experience; these bots also produce massive hate messages in the live chat within a very short time frame, making the moderator too overwhelmed to remove these accounts/messages in time [11, 25]. In this case, AI is intentionally used to perform new online attacks in an interactive and immersive online space (i.e., live streaming) 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). This type of AI-powered hate may even exacerbate harm if applied to a context like social VR, in which embodied bots with hateful messages could attack a victim in virtual space.

New User- Or Developer-Generated Harmful Design to Manipulate People's Actions. Additionally, with the increasingly popular trend to support and promote user-generated virtual worlds and immersive experiences (e.g., Roblox's business model), there is a growing concern about how these user-generated virtual worlds can be extremely harmful and manipulative, which are also hard to moderate [29]. For instance, research has shown that Roblox, a game platform primarily used by child players, allows for several patterns to design virtual worlds that harm players', especially children's, online and even offline experiences, ranging from microtransaction design that causes financial harm and social interaction design that encourages inappropriate interpersonal interactions to virtual world design that promotes harmful ideologies [29]. Likewise, immersive technology is generating new unique types of threats that are grounded in its inherent ability to control the visual perception of the user [24]. Tseng et al. [50] and Bonnail et al. [10] demonstrated how developers or content creates might leverage known perceptual manipulation techniques (e.g., redirected walking [36]) to negatively impact the memory or even physical actions of a VR user.

Therefore, we believe that seeking novel approaches to understand and mitigate these emerging and understudied new forms of online harm in immersive and embodied social spaces in the broader sense (including but not limited to XR, gaming and virtual worlds, live streaming, and video-conferencing) is a critically needed HCI research agenda for achieving a safer online environment in the future.

2 GOALS OF THE WORKSHOP

In recognition of these new and understudied harms in immersive and embodied online spaces, there is an emerging research agenda in HCI that seeks to understand and mitigate such harms in various ways. For example, focusing on embodied harassment in social VR, several design directions for new harassment mitigation mechanisms for social VR spaces have been proposed, including various platform-embedded new safety features such as voice modulators for users to conceal their gender, sexual, and racial identity [20], better reporting and documentation mechanisms [20], and consent mechanics to address interpersonal harm [42, 57]. In particular, these works have begun to explore how advanced Artificial Intelligence (AI) technologies can be leveraged to mitigate said harassment in social VR, including through AI-based moderation systems [41], AI moderator avatars for protecting children [16], and Non-Player Characters (NPCs) to educate social VR users about harassment mitigation [56]. Likewise, research on AI-powered hate raids on live streaming has warned that existing moderation tools are insufficient to stop massive bot accounts and bot-generated hate messages and can even be abused to fuel AI-powered online attacks [11]. Therefore, these works have suggested designing several new tools to particularly mitigate human-bot coordinated hate raids, such as using non-text-based communication and crowdsourcing

moderation with up and down votes [11]. Additionally, researchers have proposed a shift from content moderation to design moderation to help better regulate user-generated virtual worlds/experiences [29]. Slater et al. [44] opened the discussion among VR researchers about how increasing realism in VR might lead to even more threats to the user and make them susceptible to perceptual manipulations (e.g., memory source confusion in VR dominantly relies on the ability of the user to leverage visual cues that are associated uniquely to one source [28])

While this small body of existing work has begun to attend to these new harms emerging in various immersive and embodied online spaces, we argue that there exists an urgent need to explore further how these new harms continue to shape the current research discourse of online safety, cybersecurity, and immersive and embodied interactions in HCI, including: how the design of immersive and embodied virtual spaces may invite or discourage these new online harms, emerging online social norms that influence perceptions of these new harms, and how these harms can be understood and experienced in various ways across different populations and communities, and what new technologies and mechanisms can be envisioned, designed, and implemented to better understand and mitigate these harms. Investigating these questions requires more cross-disciplinary, community-wide discussion and collective reflections. CHI is the place where interdisciplinarity flourishes, and our goal of the workshop is to gather researchers and practitioners from various domains, including online safety, policy-making, AI development, cybersecurity, games and virtual worlds, and XR, to exchange knowledge on unpacking challenges to understand various forms of new online harms in immersive and embodied virtual spaces as well as identifying potential opportunities and limitations of leveraging new technologies such as AI to mitigate such harms. We will solicit research from all approaches and disciplines to cover topics such as:

- Identifying various forms of new online harms across different immersive and embodied virtual spaces
- Creating frameworks, taxonomies, and definitions to describe these new online harms
- Explaining potential sociotechnical causes of these new online harms
- Investigating how harms in immersive and embodied spaces can also be translated to physical harms in the offline world
- Understanding various stakeholders' perspectives when encountering these new harms in immersive and embodied virtual spaces
- Unpacking how these new online harms may especially damage marginalized technology users' online experiences
- Building a fundamental understanding of how metrics of immersive online experiences, such as presence and embodiment, that are traditionally considered as positive might amplify the negative impact of these new online harms
- Synthesizing existing strategies and recommendations to deal with these new online harms
- Articulating coping approaches for managing exposure to harms that can be integrated into system designs
- Evaluating how, if at all, existing harm mitigation strategies (e.g., moderation) can or cannot be used to mitigate these new harms in immersive and embodied online spaces
- Identifying both opportunities and risks of leveraging AI to detect and mitigate these new online harms
- Describing how AI can be used to create new and more severe forms of online harm in immersive and embodied virtual spaces rather than mitigating harms
- Brainstorming alternative novel approaches to understand and mitigate new online harms in immersive and embodied online spaces

As experts in human computer interaction, we hope to build a community of researchers interested in taking a proactive and pioneering approach to collectively discuss and reflect upon how we can leverage new approaches, technologies, and mechanisms to better understand and mitigate intensified and potentially more severe forms of new online harms in immersive and embodied virtual spaces to protect online users, especially those who are often considered marginalized and vulnerable, before such harms start to plague our future online social spaces.

3 ORGANIZERS

Our team of organizers represents experts in the area of games, live streaming, XR, and social VR with diverse research approaches (e.g., qualitative, experimental, design, and machine learning). Each brings unique expertise and vision to the topic.

Guo Freeman (main contact) is an Associate Professor of Human-Centered Computing at Clemson University. Her work focuses on how interactive technologies such as multiplayer online games, esports, live streaming, and social VR shape interpersonal relationships and group behavior; and how to design safe, inclusive, and supportive social VR spaces to combat emergent harassment risks especially for marginalized users.

Julian Frommel is an Assistant Professor in Interaction/ Multimedia at Utrecht University. He is interested in the design and implementation of interactive digital systems that provide enjoyable, meaningful, safe, and healthy experiences for users, including research on how to mitigate negative effects of toxicity and harassment in online games and other online spaces.

Regan Mandryk is a Professor of Computer Science at the University of Victoria, Canada. Her work focuses on how people of all ages use playful technologies for social, cognitive, and emotional wellbeing, how toxicity, discrimination, and harassment thwart the connection and recovery benefits provided by multiplayer games, and how we can design playful collaborative systems that benefit—not harm—player wellbeing.

Jan Gugenheimer is an Assistant Professor at TU-Darmstadt and Telecom-Paris, working on HCI-related topics in the field of Extended Reality. His focus is on understanding how new unique properties of immersive technologies can be used to deceive and manipulate a user's actions and beliefs and how we have to design the technology to prevent such potential misuse.

Lingyuan Li is a Research Associate of Human-Centered Computing at Clemson University. She delves into the intricacies of mediated experiences shaped by emerging technologies such as social VR, digital peer-to-peer payments, esports, and live streaming,

and how we can design more inclusive, safer, and supportive spaces within social VR.

Daniel Johnson is a Professor of Computer Science at Queensland University of Technology. His work focuses on how videogames influence wellbeing, often through the lenses of Self-Determination Theory and the Dualistic Model of Passion. His current focus includes better understanding and minimising toxic and disruptive behaviour in online settings, including with children.

4 WEBSITE AND PRE-WORKSHOP PLANS

We will create a website to advertise the CFP, provide details about the workshop and submissions, and introduce the organizers. Leading up to the workshop, we will add the list of accepted workshop submissions with PDFs, a schedule of workshop activities, and information related to accessibility accommodation. Following the workshop, we anticipate continuing to use the website for community building by hosting outcomes of the workshops, white papers, plans to publish the workshop papers, and special calls for participation in events related to the topic.

Prior to the workshop, we will make this workshop broadly visible in our community and further increase the awareness of the critical need for exploring novel approaches to understand and mitigate new online harms in immersive and embodied virtual spaces. We will especially focus on including participants who represent diverse backgrounds and reach out to our connections in communities that focus on working with specific populations of underrepresented individuals. We will also reach out via mailing lists and other communities on Slack and Discord, not only associated with communities and universities in North America and Europe, but also via global networks. Additionally, we will point potential attendees to SIGCHI resources on financial support for attendee participation (e.g., Gary Marsden Travel Awards) in the CHI conference.

5 HYBRID PLANS AND ASYNCHRONOUS ENGAGEMENT

We are planning a hybrid workshop. Because of the interactive nature of the planned activities, we feel that attendees would benefit from in-person attendance over an online-only format. However, due to the various considerations of reduced travel (e.g., sustainability, financial or visa restrictions, and ongoing travel restrictions due to pandemic and global mobility issues), we support participation from those unable or unwilling to travel. Plenary sessions and breakout sessions will be supported using videoconferencing tools (e.g., Zoom). Depending on the ratios of in-person and remote attendees, we will form remote breakout groups or integrate participants into hybrid groups during the interactive workshop activities, a format that we successfully implemented in our prior workshop at CHI 2023 [31]. We will ask about accessibility requirements long before the conference, arrange auto-captioning (which benefits everyone), and also arrange live captioning or an interpreter, if requested.

For those unable to participate synchronously, the website will be a repository of submissions of the workshop. Further, we will employ a Discord or Slack group that participants can use to chat leading up to the workshop, on the day, and following the event.

6 WORKSHOP ACTIVITIES

The workshop will contain four main themes of activities, demarcated by coffee breaks and lunch.

Theme One: Building Community. At the beginning of the workshop, the organizers will communicate the goals for the day. Participants and organizers will also introduce themselves to situate the knowledge that people are offering and expect to gain throughout this workshop. *The ultimate goal of this theme is to foster a sense of community and for participants and organizers to get to know each other, understand why we are interested in this problem space, and comprehend our shared group knowledge, concerns, and approaches.*

Theme Two: Collective Reflections. After participants and organizers start to have a sense of community and build mutual understanding of each other's work, we will engage in collective reflections on what we know and what we do not know (yet) about how to better understand and mitigate various new forms of online harms in immersive and embodied virtual spaces. Participants will share what they know about and do not know about these new online harms in their own research, such as in light of (1) the complicated nature and various forms of such harm across various immersive and embodied virtual spaces (e.g., live streaming, video conferencing, and XR); (2) how various populations, especially marginalized technology users (e.g., women, racial minorities, LGBTQ+ individuals, children, older adults, and disabled users), experience and manage such harms; and (3) existing and envisioned approaches (e.g., leveraging AI and consent mechanics) to detect, moderate, and prevent such harms and limitations of these approaches, among others. We will use this exercise to form breakout groups of participants interested in similar issues (e.g., understanding the phenomenon itself; understanding different populations' and communities' unique experiences; and harm mitigation strategies) but from differing perspectives (e.g., policymaking, psychology, sociology, education, computer science and engineering). Groups will be formed prior to the lunch break to encourage people to share their lunch break with new contacts. *The ultimate goal of this theme is to establish collective reflections on our existing interdisciplinary efforts to understand and mitigate emerging new online harms in immersive and embodied virtual spaces for diverse user groups and identify the challenges that face us—both as a community and individually, which will orient Theme 3 workshop activities.*

Theme Three: Black Mirror Writers' Room. Built upon our collective reflections on both opportunities and risks of existing or proposed approaches to detect and mitigate various forms of emerging online harms in immersive and embodied virtual spaces, we will then work in our breakout groups to brainstorm alternative novel approaches to understand and mitigate such harms in new and more inclusive ways. In particular, we will use a structured exercise, "Black Mirror Writers' Room" [18], which has been successfully used for teaching and discussing technology ethics through speculation. In this exercise, participants will work as breakout groups, and each take one issue regarding new online harms in immersive and embodied virtual spaces and take a step further: What if this issue escalates to a point that would be worthy of a Black Mirror episode? And what can we do to prevent this from

happening? These speculations may lead to valuable conversations about different stakeholders, responsibilities, policies and regulations, and novel technological solutions. One member will take notes for asynchronous participants, and each group will report back to the plenary session on their discussion.

Theme Four: Steps Toward the Future. We will wrap up by summarizing the day and defining future directions. We will use tools such as the Stop-Start-Continue exercise to map out how our community should move forward in both research and practice. As a community, what activities need to stop? What needs to start? What should we continue to do? Finally, we will discuss post-workshop plans for continued community building, progress, and opportunities to work together (see section 8). *The ultimate goal of this theme is to collaboratively speculate on future novel technologies, approaches, and solutions as a proactive response to the growing online risks in the increasingly popular immersive and embodied virtual spaces and to articulate tangible steps on moving forward in establishing our community concerned with people's online safety in these spaces.*

7 ACCESSIBILITY

We are firmly committed to accessibility and will ensure that all posted submissions are accessible, including alt text of figures. See section 5 for a description of how we plan to accommodate accessibility needs in the workshop itself.

8 PLANS TO PUBLISH WORKSHOP PROCEEDINGS

Our workshop materials will be hosted on our conference website, providing a repository for the discussion. The organizers will also collaboratively author a summary that will be pitched to Interactions magazine for publication or will be posted to Medium.com and cross-listed to SIGCHI. The organizers of this workshop have previously taken both of these approaches, and the outcome will depend on what types of discussion unfold at the workshop itself. We will also plan to publish the content participants submitted, such as publishing a collection of the submitted papers as workshop proceedings via https://ceur-ws.org or arXiv.

The longer-term plan is to publish a special issue of a journal or an edited book of contributions. We will bring options that we have researched in advance to the participants. At the end of the workshop, we will gather options and begin to develop the call for participation. Contribution to a special issue or edited book would not be limited to workshop participants, and the workshop organizers would not be the de facto editors. Rather, the workshop would act as a catalyst for one or more research collections, including the editorial team. In addition to research outputs, we also aim for the workshop to motivate future events (e.g., panels/workshops at other SIGCHI venues), research collaborations, and grant applications.

9 CALL FOR PARTICIPATION

Immersive and embodied virtual spaces are increasingly prevalent. Such environments are also generally beneficial, facilitating connection with others and allowing for novel experiences. However, the paradigm shifts introduced by such environments (e.g., immersion, embodiment, or technological advancements) can also lead to new harms for users, such as *embodied harassment* in social VR [20], *racist Zoombombing* [34], human-bot coordinated *hate raids* in live streaming [11], harmful user-generated virtual world design [29], and new types of perceptual manipulations leveraging the core abilities of XR technology to create illusions to the user [10, 24, 50].

In this workshop, we want to bring together a set of interdisciplinary researchers and practitioners from HCI and adjacent fields to better understand these new harms and develop novel strategies to mitigate them. We invite interested researchers and practitioners to submit a short position statement (2 pages maximum in CHI submission format) via the workshop website: https://sites.google.com/view/chi-2024-workshop-newharms/, in which they describe why they are interested in the area, which work they have already done, and if they are interested in specific topics related to the workshop, including but not limited to:

- Identifying various forms of new online harms across different immersive and embodied virtual spaces
- Creating frameworks, taxonomies, and definitions to describe these new online harms
- Explaining potential sociotechnical causes of these new online harms
- Investigating how harms in immersive and embodied spaces can also be translated to physical harms in the offline world
- Understanding various stakeholders' perspectives when encountering these new harms in immersive and embodied virtual spaces
- Unpacking how these new online harms may especially damage marginalized technology users' online experiences
- Building a fundamental understanding of how traditionally considered positive metrics of immersive online experiences such as presence and embodiment might amplify the negative impact of these new online harms
- Synthesizing existing strategies and recommendations to deal with these new online harms
- Articulating coping approaches for managing exposure to harms that can be integrated into system designs
- Evaluating how, if at all, existing harm mitigation strategies (e.g., moderation) can or cannot be used to mitigate these new harms in immersive and embodied online spaces
- Identifying both opportunities and risks of leveraging AI to detect and mitigate these new online harms
- Describing how AI can be used to create new and more severe forms of online harm in immersive and embodied virtual spaces rather than mitigating harms
- Brainstorming alternative novel approaches to understand and mitigate new online harms in immersive and embodied online spaces

We will accept submissions in the scope of the workshop in which participants have prior expertise. The workshop will be organized as a hybrid event at CHI 2024. One participant from each submission must register for the workshop and at least one day of the CHI conference. All submissions will be published on the workshop website and a collection of workshop papers.

REFERENCES

[1] Veronica Abebe, Gagik Amaryan, Marina Beshai, Ilene, Ali Ekin Gurgen, Wendy Ho, Naaji R Hylton, Daniel Kim, Christy Lee, Carina Lewandowski, et al. 2022. Anti-Racist HCI: notes on an emerging critical technical practice. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–12.

[2] Julia Alexander. 2018. 'Ugandan Knuckles' is overtaking VRChat. https://www.polygon.com/2018/1/8/16863932/ugandan-knuckles-meme-vrchat

[3] Anti-Defamation League. 2021. Hate is no game: Harassment and positive social experiences in online games 2021. https://www.adl.org/hateisnogame#executive-summary

[4] Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. 2021. CyberPolice: Classification of Cyber Sexual Harassment. In *EPIA Conference on Artificial Intelligence*. Springer, 701–714.

[5] Nicole A Beres, 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 (CHI '21)*. Association for Computing Machinery, Yokohama, Japan, 1–15. https://doi.org/10.1145/3411764.3445157

[6] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.

[7] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[8] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social VR: Implications for design. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 854–855.

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

[10] Elise Bonnail, Wen-Jie Tseng, Mark Mcgill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2023. Memory Manipulations in Extended Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 875, 20 pages. https://doi.org/10.1145/3544548.3580988

[11] 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).

[12] 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.

[13] Julian Dibbell. 1993. A Rape in Cyberspace: or, How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database Into a Society. http://www.juliandibbell.com/texts/bungle_vv.html

[14] 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.

[15] Bryan Dosono, Ihudiya Finda Ogbonnaya-Ogburu, Yolanda A Rankin, Angela DR Smith, and Kentaro Toyama. 2022. Anti-Racism in Action: A Speculative Design Approach to Reimagining SIGCHI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–5.

[16] Cristina Fiani, Robin Bretin, Mark McGill, and Mohamed Khamis. 2023. Big Buddy: Exploring Child Reactions and Parental Perceptions towards a Simulated Embodied Moderating System for Social Virtual Reality. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 1–13.

[17] Cristina Fiani and Stacy Marsella. 2022. Investigating the Non-Verbal Behavior Features of Bullying for the Development of an Automatic Recognition System in Social Virtual Reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*. 1–3.

[18] Casey Fiesler. 2018. Black Mirror, Light Mirror: Teaching Technology Ethics Through Speculation. https://cfiesler.medium.com/the-black-mirror-writers-room-teaching-technology-ethics-through-speculation-f1a9e2deccf4.

[19] Guo Freeman and Divine Maloney. 2021. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[20] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30.

[21] Julian Frommel, Daniel Johnson, and Regan L Mandryk. 2023. How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness. *Computers in Human Behavior Reports* 10 (2023), 100302.

[22] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. " You have to prove the threat is real": Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[23] Kishonna L Gray. 2012. Deviant bodies, stigmatized identities, and racist acts: Examining the experiences of African-American gamers in Xbox Live. *New Review of Hypermedia and Multimedia* 18, 4 (2012), 261–276.

[24] Jan Gugenheimer, Wen-Jie Tseng, Abraham Hani Mhaidli, Jan Ole Rixen, Mark McGill, Michael Nebeling, Mohamed Khamis, Florian Schaub, and Sanchari Das. 2022. Novel Challenges of Safety, Security and Privacy in Extended Reality. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. https://doi.org/10.1145/3491101.3503741

[25] 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.

[26] Qinglai He, Yili Kevin Hong, and TS Raghu. 2022. The Effects of Machine-powered Content Modereation: An Empirical Study on Reddit. In *55th Hawaii International Conference on System Sciences (HICSS)*.

[27] 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.

[28] Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. 1993. Source monitoring. *Psychological bulletin* 114, 1 (1993), 3.

[29] Yubo Kou and Xinning Gui. 2023. Harmful Design in the Metaverse and How to Mitigate It: A Case Study of User-Generated Virtual Worlds on Roblox. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 175–188. https://doi.org/10.1145/3563657.3595960

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

[31] Regan L Mandryk, Julian Frommel, Nitesh Goyal, Guo Freeman, Cliff Lampe, Sarah Vieweg, and Donghee Yvette Wohn. 2023. Combating Toxicity, Harassment, and Abuse in Online Social Spaces: A Workshop at CHI 2023. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[32] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping pro-social interaction in VR: an emerging design framework. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[33] Maria D Molina and S Shyam Sundar. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society* (2022), 14614448221103534.

[34] Lisa Nakamura, Hanah Stiverson, and Kyle Lindsey. 2021. *Racist Zoombombing*. Routledge.

[35] Jessica Outlaw and Beth Duckles. 2018. Virtual Harassment: The Social Experience of 600+ Regular Virtual Reality (VR) Users. https://virtualrealitypop.com/virtual-harassment-the-social-experience-of-600-regular-virtual-reality-vr-users-23b1b4ef884e

[36] Sharif Razzaque. 2005. *Redirected walking*. The University of North Carolina at Chapel Hill.

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

[38] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, Vol. 2. IEEE, 241–244.

[39] Nazanin Sabri, Bella Chen, Annabelle Teoh, Steven P Dow, Kristen Vaccaro, and Mai Elsherief. 2023. Challenges of Moderating Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[40] Kelsea Schulenberg, Guo Freeman, Lingyuan Li, and Catherine Barwulor. 2023. Creepy Towards My Avatar Body, Creepy Towards My Body: How Women Experience and Manage Harassment Risks in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2023). https://guof.people.clemson.edu/papers/cscw23women.pdf

[41] 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*.

[42] Kelsea Schulenberg, Lingyuan Li, Caitlin Lancaster, Doug Zytko, and Guo Freeman. 2023. "We Don't Want a Bird Cage, We Want Guardrails"': Understanding & Designing for Preventing Interpersonal Harm in Social VR through the Lens of Consent. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2023). https://guof.people.clemson.edu/papers/cscw23consent.pdf

[43] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*. IEEE, 225–226.

[44] Mel Slater, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, et al. 2020. The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality* 1 (2020), 1.

[45] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria V Sanchez-Vives. 2009. Inducing illusory ownership of a virtual body. *Frontiers in neuroscience*

(2009), 29.

[46] Weilun Soon. 2022. A researcher's avatar was sexually assaulted on a metaverse platform owned by Meta. https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5

[47] Hannah Sparks. 2021. Woman claims she was virtually 'groped' in Meta's VR metaverse. https://nypost.com/2021/12/17/woman-claims-she-was-virtually-groped-in-meta-vr-metaverse/

[48] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. 2020. " They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–29.

[49] Stephanie Tom Tong, Elizabeth Stoycheff, and Rahul Mitra. 2022. Racism and resilience of pandemic proportions: online harassment of Asian Americans during COVID-19. *Journal of Applied Communication Research* (2022), 1–18.

[50] Wen-Jie Tseng, Elise Bonnail, Mark McGill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2022. The Dark Side of Perceptual Manipulations in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 612, 15 pages. https://doi.org/10.1145/3491102.3517728

[51] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.

[52] Leijie Wang and Haiyi Zhu. 2022. How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 824–838.

[53] Sai Wang. 2021. Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital journalism* 9, 1 (2021), 64–83.

[54] Michel Wijkstra, Katja Rogers, Regan L Mandryk, Remco C Veltkamp, and Julian Frommel. 2023. Help, My Game Is Toxic! First Insights from a Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 3–9.

[55] Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7, 3 (1998), 225–240.

[56] Qingxiao Zheng, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan C Salvi, Guo Freeman, and Yun Huang. 2023. Understanding Safety Risks and Safety Design in Social VR Environments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–37.

[57] Douglas Zytko and Jonathan Chan. 2023. The Dating Metaverse: Why We Need to Design for Consent in Social VR. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2489–2498.