Combating Toxicity, Harassment, and Abuse in Online Social Spaces: A Workshop at CHI 2023

Regan L. Mandryk
regan@acm.org
University of Saskatchewan
Canada

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

Nitesh Goyal
niteshgoyal@acm.org
Google
USA

Guo Freeman
guof@clemson.edu
Clemson University
USA

Cliff Lampe
cacl@umich.edu
University of Michigan
USA

Sarah Vieweg
sarahvieweg@gmail.com
Cash App
USA

Yvette Wohn
wohn@njit.edu
New Jersey Institute of Technology
USA

ABSTRACT

Online social spaces provide much needed connection and belonging—particularly in a context of continued lack of global mobility due to the ongoing Covid-19 pandemic and climate crisis. However, the norms of online social spaces can create environments in which toxic behaviour is normalized, tolerated or even celebrated. This can occur without consequence, leaving its members vulnerable to hate, harassment, and abuse. A vast majority of adults have experienced toxicity online and the harm is even more prevalent for members of marginalized and minoritized groups, who are more often the targets of online abuse. Although there is significant work on toxicity in the SIGCHI community, approaches and knowledge have typically been siloed by the domain of investigation (e.g., social media, multiplayer games, social VR). We argue that cross-disciplinary efforts will benefit not only the various communities and situations in which abuse occurs, but that bringing together researchers from different backgrounds and specialties will provide a robust and rich understanding of how to tackle online toxicity at scale.

CCS CONCEPTS

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

KEYWORDS

toxicity, harassment, abuse, hate speech, flaming, trolling, griefing, doxxing, social media, multiplayer games, social VR, metaverse

ACM Reference Format:
Regan L. Mandryk, Julian Frommel, Nitesh Goyal, Guo Freeman, Cliff Lampe, Sarah Vieweg, and 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 (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3573793

1 BACKGROUND

Trigger Warning: In this workshop description, we discuss toxicity expressed online in social spaces, including harassment, hate, abuse, rape, groping, flaming, trolling, griefing, and doxxing.

People are increasingly socializing through a wide range of technologies: we video chat using *Zoom, Google Meet etc.* with our therapists, employers, and students, and use *WhatsApp, Chat, Message etc.* with our friends, while using *Slack, Google Meet, Teams etc.* with our colleagues. We form and maintain relationships by playing multiplayer games and engaging with social media, and we can even meet our life partners in an app or through embodied avatars in social Virtual Reality (VR). Research consistently demonstrates that engaging with others in online social spaces can facilitate connectedness and belonging that builds social capital (e.g., [14, 15, 37]), provides emotional (e.g., listening, comforting, venting) and instrumental (e.g., helping, providing tangible aid) social support (e.g., [22, 43, 48, 52]), and combats loneliness (e.g., [12, 37, 53]). However, research also consistently demonstrates that in places where social interactions occur, people are exposed to the toxic behaviors of others, either as a bystander or as a target, making combating toxicity one of the biggest challenges for popular online social spaces. For example, the Anti-defamation League (ADL) reports that 40% of American adults have experienced harassment on social media sites [35], and this number rises to 65% among people from a marginalized or minoritized identity group. In gaming contexts, the situation is even worse: ADL reports that 83% of American adult gamers have experienced harassment in online multiplayer games [34] with consistent increases over time in identity-based harassment for women, players of colour, and LGBTQ+ players [ibid]. Beyond the commercially successful online social spaces of social media and multiplayer gaming, nascent social technologies,

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573793

such as Social VR, are being used to facilitate more nuanced online interactions through immersive and embodied experiences, but these emerging social spaces have already been identified as exposing users to more physicalized and embodied forms of harassment and other toxic behaviours (e.g., virtual "groping" and rape) [4, 23, 45, 46].

Toxicity broadly refers to various types of negative behaviors involving abusive communications directed towards other members of the online space (e.g., harassment, verbal abuse, hate speech, and flaming) and disruptive behaviours that violate the rules and social norms of the online space (e.g., spamming, trolling) [34, 35, 49]. Further, toxicity in online spaces can extend beyond the digital realm, as harassment extends into the material world (e.g., doxxing [34, 35], swatting [34, 35]). Although toxicity is a pressing problem in 2022, these behaviours have been consistently present over the history of online social interactions: for example, a rape within the multiplayer gaming space Roblox in 2018 prompted the Village Voice to republish an article on cyber-rape first published in 1993 [13]. When online spaces are disrupted by toxic behaviours, there are harmful effects that extend to all stakeholders. Toxicity has also been defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion"[1], causing companies that host these online spaces to experience harm from user disengagement and churn (e.g., [28, 33]). From the user perspective, experiencing toxicity can harm mood and enjoyment (e.g, [49]), can affect people's ability to execute tasks (e.g., [39]), and can cause lasting distress, similar to experiencing toxic interactions in the material world (e.g., [4]). The harms are particularly relevant for marginalized and minoritized groups (e.g., women, people of colour, LGBTQ+ people), who are more likely to be targets of toxic behaviour [34, 35] and for whom the lasting harms extend into health and wellness [3]. Even identifying toxicity at scale and responsibly using AI/ML models can be inequitable to some identity groups. Recent research has shown that attacks on certain communities, e.g. LGBTQ+ and African Americans, might not be appropriately identified by AI owing to the quality of training data provided by data annotators since they have limited lived experiences, knowledge and participation in such communities [26]. This is clearly a challenge that needs to be addressed from a Responsible AI perspective, especially if we need a solution that scales well. Finally, the online communities themselves are at risk from toxic behaviours as these types of harmful interactions become normalized, accepted, and considered as the status quo of the community [2, 9].

1.1 Toxicity in Gaming and Live Streaming

Multiplayer games are infamous for their toxic communities. While this is not necessarily the case for all games, it is true that toxicity, harassment, and hate are widespread in many online multiplayer games. Researchers have investigated disruptive play behaviours such as "griefing" [17] and "trolling" [11] for many years. Similarly, industry has recognized the problem and is working on solutions, e.g., by implementing detection approaches in their own games [5, 10, 44] and more at-large by contributing to organizations like the Fair Play Alliance (*www.fairplayalliance.org*). Despite this, harmful behaviours like harassment and hate are increasingly widespread in gaming, as exemplified by the aforementioned ADL report suggesting that 83% of adult gamers have experienced harassment in online multiplayer games [34]. Similar to other spaces, these behaviours are often identity-based, disproportionately targeting women, players of colour, and LGBTQ+ players [1, 19, 29, 34]. Gaming is a particularly challenging context in which to combat toxicity. For example, there is an increasing normalization of toxicity in gaming communities [2, 18], in which it is accepted that toxicity is part of gaming culture. Further, culture and language in gaming complicate the implementation of solutions, for example because of words like 'kill' that may be innocuous in games (e.g., "You already have four kills! Clutch game!"), or may be a marker of hate speech. Many of these problems extend beyond games themselves, e.g., to game streaming platforms that have to deal with challenges like "hate raids" and the rise of extremism [30]. Overall, toxicity, harassment, and hate represent a threat to the health of players and all members of gaming communities. We have to understand and combat those behaviours to ensure that multiplayer games are safe spaces for everyone.

1.2 Toxicity in Social Media

While toxicity is experienced in gaming, it is ubiquitous across multiple social media platforms and commenting sections of multiple Internet forums or communities [24]. Unfortunately what starts online does indeed transform into offline harassment in the physical spaces as well [8]. Hence, it is imperative to manage this toxicity as soon as it appears online, as the targets of online toxic behavior can face dangerous repercussions. Multiple platforms have been pursuing use of technology to scale moderation [25, 31], including developing machine learning (ML) based tools like Perspective API to detect toxicity probabilistic distribution in the text[1]. To help targets of online hate and harassment, multiple prototypical tools that leverage artificial intelligence (AI) [27, 32], moderators/screenshots [7, 47, 51], and support networks [16, 36] have been designed. Despite progress and work by multiple researchers and industry partners, multiple studies continue to point out that certain members of the society owing to their identity—e.g., women, LGBTQ+, people of color, and younger individuals—continue to be more likely to be targets of online harassment [3, 50]. Further, while some members of the community have a less than ideal option to just turn off their online presence, and reduce online consumption; such opportunity is unavailable to folks like journalists and activists who need to remain online and engage with the wider community to seek justice and share their truth, as well as content creators and other entrepreneurs whose livelihood is dependent on having an online presence (e.g.,[6, 42, 50]). More work is needed to help such targets of online harassment manage toxicity directed at them and to reduce overall online toxicity production.

1.3 Toxicity in Social VR and the Metaverse

In social VR, multiple users can interact with one another through VR head-mounted displays and immersive 360 degree virtual content in 3D virtual spaces [23, 38]. Popular social VR platforms, such as Meta's recently launched Metaverse paradigm, are increasingly innovating online interactions by providing full-body tracked avatars, predominant voice communication, body language and

[1] www.perspectiveapi.com

gestures, and simulated immersive activities. However, there is a growing concern regarding how social VR's novel features may also lead to new and more severe forms of harassment compared to other online contexts. These incidents have been frequently described in mass media and technology reports, such as trash talking/drawing penises [40, 41], the virtual "groping" behaviors, and the most recent "rape" in the metaverse [13, 45, 46]. Prior work in HCI has identified several new characteristics of online harassment in social VR and highlighted embodied harassment as an emerging but understudied form of harassment in novel online social spaces [4, 23]. These works have also pointed out that social VR users who are already considered marginalized in the gaming and virtual worlds contexts (e.g., women, LGBTQ, and ethnic minorities) may also experience such harassment in a more disruptive way because they are more identifiable, feel more vulnerable, and face a higher risk of stalking even out of social VR [20, 21, 23, 40, 41]. Understanding and mitigating emergent forms of toxicity and harassment in social VR spaces and the evolving metaverse paradigm thus is a critically needed research agenda for achieving a safer online environment.

1.4 Summary

Engaging with others online through social media, multiplayer games, or in the metaverse can facilitate connectedness and belonging, expand our social networks, and help us form and maintain relationships. However, research consistently demonstrates that people are also exposed to the toxic behaviors of others, including harmful communications (e.g., hate speech, verbal abuse), disruptive behaviours (e.g., griefing, trolling), embodied harassment (e.g., virtual groping), and physical harm that extends into offline spaces (e.g., doxxing, swotting). A vast majority of adults have experienced toxicity online, and across platforms and domains, the problem is even more prevalent for members of marginalized and minoritized groups, who are more often the targets of online abuse. Combating toxicity in all of these spaces is necessary so that people can enjoy safe interactions online.

1.5 Goals of the Workshop

Although there is significant work on toxicity in the SIGCHI community, approaches and knowledge have typically been siloed by the domain of investigation. Research on toxicity in gaming tends to be published in game-specific venues (e.g., CHI PLAY) whereas work on toxicity in social media tends to be published in venues specific to collaboration (e.g., CSCW) or design-centered venues for tools designed to manage toxicity (e.g., CHI) . Even not-for-profit groups—such as the ADL—survey people based on their experiences of hate and harassment on social media [35] or within games [34]. We argue that cross-disciplinary collaborations benefit not only the various communities and situations in which abuse occurs, but bringing together researchers from different backgrounds and specialties will provide a robust and rich understanding of how to tackle these problems at scale. Knowledge exchange among domains is particularly important as emerging technologies that blend characteristics from gaming, traditional media, and social media—such as social VR or live streaming—continue to gain popularity. CHI is the place where interdisciplinarity flourishes, and our goal on the day of the workshop is to gather researchers and practitioners from various domains to exchange knowledge on combating toxicity in online social spaces, advancing both research and practice. We will solicit research from all approaches and disciplines to cover topics such as:

- Creating frameworks, taxonomies, and definitions of online toxicity
- Understanding motivations for toxic behaviour online
- Identifying the cycle of toxicity and the emergence of new toxic behaviors online
- Characterizing how toxicity is normalized in online communities
- Detecting online toxic behaviours and communications
- Synthesizing existing strategies and mechanisms to mitigate toxic behaviors online
- Developing interventions to better combat online toxicity
- Evaluating the efficacy of interventions that combat online toxicity
- Implementing support for the targets of online toxicity
- Fostering an inclusive culture to support safer online communities
- Identifying needs of platform providers
- Supporting community figures who are fighting toxic cultures
- Educating about the current state and dangers of harmful behaviours
- Bringing members of academia and industry together to foster safer communities
- Understanding what role can AI play in identifying and preventing online toxicity

Online social spaces provide much needed connection and belonging—particularly in a context of continued lack of global mobility due to the ongoing Covid-19 pandemic and climate crisis. However, any spaces where people gather leave its members vulnerable to toxicity, hate, harassment, and abuse. Further, the community norms and anonymity of online social spaces can create environments in which toxic behaviour is normalized, tolerated or even celebrated, and occurs without consequences for the perpetrators. As experts in human computer interaction, we have a responsibility to investigate online toxicity and provide knowledge to researchers and practitioners that will aid in combating toxicity. With this workshop, we hope to build a community of experts interested in making online social spaces safer, so that everyone can benefit from innovations in social technologies that bring us together.

2 ORGANIZERS

Our team of organizers represents a range of domains (e.g., games, social media, live streaming, and social VR), research approaches (e.g., qualitative, experimental, design, machine learning), disciplines (e.g., computer science, communications), and perspectives (e.g., academia, industry). Each brings unique expertise and vision to the topic.

Regan Mandryk (main contact) is a Canada Research Chair in Digital Gaming Technologies and Experiences and Professor of Computer Science at the University of Saskatchewan. Her work focuses on how people use playful technologies for social and emotional wellbeing, and how toxicity thwarts the connection and recovery benefits provided by multiplayer games.

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 digital spaces.

Nitesh Goyal leads research on tools designed to build AI responsibly at Google Research. His work has focussed on AI for social good for marginalized populations, including creating tools for journalists/activists to manage online harassment, tools that reduce biases during investigative sensemaking, and unpacking the role of data annotators' identity that power the AI behind these tools and more.

Guo Freeman is an Assistant 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.

Cliff Lampe is a Professor and Associate Dean in the University of Michigan School of Information. His work examines how the design of systems shape and are shaped by social processes like harassment, toxicity and extremism. That work has been focused on social media and collaborative creation sites.

Sarah Vieweg is currently a Staff Researcher at Cash App, formerly at Meta and Twitter. Her experience includes work on human rights issues, abuse and harassment, misinformation, violence among young people, child safety, and issues of trust in social media and fintech spaces.

Yvette Wohn is an associate professor at New Jersey Institute of Technology. Her research uses a social support lens to examine how volunteer moderators and content creators work together in different types of online communities to proactively and reactively deal with online harassment and other forms of toxicity. Her work has led to design of new moderation tools and development of new hate speech detection algorithms.

3 WEBSITE

Our website (https://combatingonlinetoxicity.sites.uu.nl/) advertises the CFP, provides details about the workshop and how to submit, introduces the organizers, and provides resources to those interested in reading more about this domain. Leading up to the workshop, we will add bios from consenting participants, the workshop submissions, a schedule of events, 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, and special calls for participation in events related to the topic.

4 PRE-WORKSHOP PLANS

Using the website, we will communicate with attendees (and the community) prior to and following the workshop. Given our focus on toxicity, which is often targeted at people from marginalized groups, the composition of participants should reflect the diversity of people affected. We will make this workshop broadly visible in our community. First, issues surrounding marginalization are fundamental to our theme; we will reach out to our connections in communities who focus on working with specific populations of underrepresented people. Second, we will reach out via mailing lists not only associated with communities and universities in North America and Europe, but also via global networks. Third, as facilitators of the workshop, we have ensured a range of expert perspectives across various axes of identity, including gender, sexuality, (dis)ability, race-ethnicity, culture, and neuroatypicality. We also include facilitators from both academic and industrial perspectives to ensure a diversity of perspective. Fourth, we will point potential attendees to SIGCHI resources on financial support for attendee participation in the CHI conference.

In addition, we will be asking each participant to create a 1–2 minute video presentation of their submission that all attendees can watch prior to the workshop itself.

5 IN-PERSON, HYBRID, OR VIRTUAL-ONLY

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 ongoing travel restrictions due to pandemic and global mobility issues, we support participation from those unable 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 working sessions. 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.

6 ASYNCHRONOUS ENGAGEMENT

For those unable to participate synchronously, the website will be a repository of submissions, video presentations, and liveblogging of the workshop. Each breakout group will update the shared notes. Further, we will employ a discord group that participants can use to chat leading up to the workshop, on the day, and following the event.

7 WORKSHOP STRUCTURE

The workshop will contain four blocks of activity, demarcated by coffee breaks and lunch.

7.1 Block One: Building Community

Goal: To get participants to know who each other are, what they bring to the table, and what our shared group knowledge is.

The initial block will be spent communicating the goals for the day, and having participants and facilitators introduce themselves to situate the knowledge that people are offering over the course of the day. Introductions will be structured and facilitated to ensure efficient knowledge translation along with two-way communication between attendees. We will complete the initial block with a quick brainstorming session to determine whether participants are focused on online toxicity in terms of foundational understanding, detection, intervention, or support.

7.2 Block Two: Taking Stock

Goal: To collaboratively take stock of where the community is at, in terms of research and practice, and identify the challenges that face us—both as a community and individually.

Organizers will kick off this session with brief presentations on the state-of-the-art in their domain (e.g., online dating, gaming). Following these we will establish as a group what we collectively know and wish we knew about online toxicity, using a gamestorming exercise[2] that also highlights our domain of expertise. Participants will share what they know about toxicity in their own domain, identify common or unique challenges for combating harassment and toxicity across various online contexts, and describe how successful examples (or failures) of combating toxicity in one context may inform other contexts. We will use this exercise to form breakout groups of participants interested in similar issues (e.g., detecting toxicity, intervention design) but from differing perspectives (e.g., social media, livestreaming). Groups will be formed prior to the lunch break to encourage people to share their lunch break with new contacts.

7.3 Block Three: Imagining Solutions

Goal: To collaboratively speculate on future technologies, approaches, and solutions.

We will work in our breakout groups on a structured exercise to brainstorm future solutions, the barriers that we face before implementing these solutions, and what steps could be taken to address these barriers. One member will take notes for asynchronous participants and each group will report back to the plenary session on their discussion.

7.4 Block Four: Making Plans

Goal: To articulate tangible steps on moving forward in establishing our community.

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 9.2).

8 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.

9 POST-WORKSHOP PLANS

9.1 Short-term Plans: Workshop Summary

Our workshop materials will be hosted on our conference website, providing a repository of 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 crosslisted to SIGCHI. The organizers of this workshop have previously taken both of these approaches, and the outcome will depend on what types of discussion unfolds at the workshop itself.

9.2 Long-term Plans: Special Issues and Future Events

The longer-term plan is to publish a special issue of a journal or an edited book of contributions. We will bring options to the participants, researched in advance, to gather opinions at the workshop (in block four) 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., Dagstuhl seminar), research collaborations, and grant applications.

10 CALL FOR PARTICIPATION

Online social spaces (e.g., social media, multiplayer games, social VR) provide much needed connection and belonging—particularly in a context of continued lack of global mobility due to the ongoing Covid-19 pandemic and climate crisis. However, the norms of online social spaces can create environments in which toxic behaviour is normalized, tolerated or even celebrated, and occurs without consequence, leaving its members vulnerable to hate, harassment, and abuse. With this workshop, we hope to build a community of experts interested in combating online toxicity.

We invite researchers and practitioners to submit short position statements (2 pages maximum in CHI submission format) via the workshop website that describe their knowledge and their domain of expertise, on topics such as:

- Creating frameworks, taxonomies, and definitions of online toxicity
- Understanding motivations for toxic behaviour
- Characterizing the emergence and normalization of toxicity
- Detecting toxicity through communication and behaviours
- Developing and evaluating interventions to combat toxicity
- Implementing support for the targets of toxicity
- Fostering inclusive cultures to support safer online communities
- Identifying needs of platform providers

We will accept submissions within scope of the workshop in which participants have demonstrated expertise. The workshop will be in person at CHI 23 with support for remote attendees. Participants will be asked to submit an introduction video (1–2 minutes) prior to the conference, and one participant from each submission must register for the workshop and at least one day

[2] www.gamestorming.com/the-blind-side/

of the CHI conference. All submissions will be published on the workshop website.

REFERENCES

[1] Mary Elizabeth Ballard and Kelly Marie Welch. 2017. Virtual warfare: Cyberbullying and cyber-victimization in MMOG play. *Games and Culture* 12, 5 (2017), 466–491.

[2] 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 CHI 2021 (CHI '21)*. ACM, Yokohama, Japan, 1–15. https://doi.org/10.1145/3411764.3445157

[3] 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.

[4] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 100 (nov 2019), 25 pages. https://doi.org/10.1145/3359202

[5] Blizzard. 2022. Defense Matric Activated! Fortifying Gameplay Integrity and Positivity in Overwatch 2. https://overwatch.blizzard.com/en-us/news/23857517/defense-matrix-activated-fortifying-gameplay-integrity-and-positivity-in-overwatch-2/

[6] Ross Bonifacio, Lee Hair, and Donghee Yvette Wohn. 2021. Beyond fans: The relational labor and communication practices of creators on Patreon. *New Media & Society* (2021), 14614448211027961.

[7] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[8] Rohit Kumar Chandaluri and Shruti Phadke. 2019. *Cross-Platform Data Collection and Analysis for Online Hate Groups*. Ph. D. Dissertation. Virginia Tech.

[9] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (nov 2018), 25 pages. https://doi.org/10.1145/3274301

[10] Joel Chapman and FACEIT. 2021. Fighting Abusive Behaviour — Product Update. https://medium.com/faceit-blog/fighting-abusive-behaviour-product-update-d39e490c00ce

[11] Christine Cook, Juliette Schaafsma, and Marjolijn Antheunis. 2017. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society* 20, 9 (Dec. 2017), 3323–3340. https://doi.org/10.1177/1461444817748578

[12] Ansgar E. Depping, Colby Johanson, and Regan L. Mandryk. 2018. Designing for Friendship: Modeling Properties of Play, In-Game Social Capital, and Psychological Well-Being. In *Proc. of CHI PLAY '18* (Melbourne, Australia). ACM, NY, USA, 87–100. https://doi.org/10.1145/3242671.3242702

[13] Julian Dibbel. 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. https://www.villagevoice.com/2018/07/25/before-roblox-an-online-rape-when-cyberspace-was-new/

[14] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x

[15] Nicole B. Ellison, Jessica Vitak, Rebecca Gray, and Cliff Lampe. 2014. Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes. *JCMC* 19, 4 (07 2014), 855–870. https://doi.org/10.1111/jcc4.12078

[16] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting online harassment of women journalists. In *Mediating misogyny*. Springer, 311–332.

[17] Chek Yang Foo and Elina M. I. Koivisto. 2004. Defining grief play in MMORPGs: player and developer perceptions. In *Proc. of ACE 2004 (ACE '04)*. Association for Computing Machinery, Singapore, 245–250. https://doi.org/10.1145/1067343.1067375

[18] Jesse Fox, Michael Gilbert, and Wai Yen Tang. 2018. Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media & Society* 20, 11 (2018), 4056–4073.

[19] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307. https://doi.org/10.1177/1461444816635778

[20] Guo Freeman and Dane Acena. 2022. " Acting Out" Queer Identity: The Embodied Visibility in Social Virtual Reality. *Proceedings of the ACM on humancomputer interaction* 6, CSCW2 (2022).

[21] Guo Freeman, Divine Maloney, Dane Acena, and Catherine Barwulor. 2022. (Re) discovering the Physical Body Online: Strategies and Challenges to Approach Non-Cisgender Identity in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[22] Guo Freeman and Donghee Yvette Wohn. 2017. Social support in eSports: building emotional and esteem support from instrumental support interactions in a highly competitive environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 435–447.

[23] 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.

[24] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

[25] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[26] Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *arXiv preprint arXiv:2205.00501* (2022).

[27] 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 *Proc. of CHI 2022* (New Orleans, LA, USA). ACM, New York, NY, USA, Article 242, 17 pages. https://doi.org/10.1145/3491102.3517517

[28] Kate Grandprey-Shores, Yilin He, Kristina L. Swanenburg, Robert Kraut, and John Riedl. 2014. The Identification of Deviance and Its Impact on Retention in a Multiplayer Game. In *Proc. of CSCW 2014* (Baltimore, USA). ACM, NY, USA, 1356–1365. https://doi.org/10.1145/2531602.2531724

[29] 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.

[30] Nathan Grayson. 2022. How Twitch took down Buffalo shooter's stream in under two minutes. https://www.washingtonpost.com/video-games/2022/05/20/twitch-buffalo-shooter-facebook-nypd-interview/

[31] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[32] Jaeheon Kim, Donghee Yvette Wohn, and Meeyoung Cha. 2022. Understanding and identifying the use of emotes in toxic chat on Twitch. *Online Social Networks and Media* 27 (2022), 100180.

[33] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* (2020).

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

[35] Anti-Defamation League. 2022. Online Hate and Harassment: The American Experience 2022. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022

[36] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[37] Regan L. Mandryk, Julian Frommel, Ashley Armstrong, and Daniel Johnson. 2020. How Passion for Playing World of Warcraft Predicts In-Game Social Capital, Loneliness, and Wellbeing. *Frontiers in Psychology* 11, September (2020). https://doi.org/10.3389/fpsyg.2020.02165

[38] 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.

[39] Charles K. Monge and T. Caitlin Vasquez-O'Brien. 2022. Effects of individual toxic behavior on team performance in League of Legends. *Media Psychology* 25, 1 (2022), 82–105. https://doi.org/10.1080/15213269.2020.1868322

[40] Jessica Outlaw and Beth Duckles. 2017. Why Women Don't Like Social Virtual Reality: A Study of Safety, Usability, and Self-expression in Social VR. https://www.extendedmind.io/why-women-dont-like-social-virtual-reality

[41] 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

[42] Andrew Phelps and Mia Consalvo. 2020. Laboring Artists: Art Streaming on the Videogame Platform Twitch. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[43] Nathaniel Poor, Marko Skoric, and Cliff Lampe. 2022. Death of a child, birth of a guild: Factors aiding the rapid formation of online support communities. *The Information Society* 38, 3 (2022), 188–199. https://doi.org/10.1080/01972243.2022.2071219

[44] Riot Games. 2022. Valorant Voice Evaluation Update. https://playvalorant.com/en-us/news/announcements/valorant-voice-evaluation-

update/?linkId=100000132203363

[45] 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

[46] 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/

[47] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, et al. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proc. of CHI 2021*. 1–18.

[48] Sabine Trepte, Leonard Reinecke, and Keno Juechems. 2012. The social side of gaming: How playing online computer games creates online and offline social support. *Computers in Human behavior* 28, 3 (2012), 832–839.

[49] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proc. of CHI 2020*. ACM, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376191

[50] Jirassaya Uttarapong, Jie Cai, and Donghee Yvette Wohn. 2021. Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity. In *ACM International Conference on Interactive Media Experiences*. 7–19.

[51] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[52] Donghee Yvette Wohn, Guo Freeman, and Caitlin McLaughlin. 2018. Explaining viewers' emotional, instrumental, and financial support provision for live streamers. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[53] Donghee Yvette Wohn and Cliff Lampe. 2018. Psychological Wellbeing as an Explanation of User Engagement in the Lifecycle of Online Community Participation. In *Proc. of GROUP 2018* (Sanibel Island, Florida, USA). ACM, NY, USA, 184–195. https://doi.org/10.1145/3148330.3148351