Hate Speech in the Metaverse

Esen K. Tütüncü & Danielle Shanley

Hate speech, as defined by the United Nations (UN), refers to any form of communication, gesture, or conduct that may incite violence, discrimination, hostility, or prejudicial action against individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, or other characteristics. It encompasses expressions that demean, dehumanize, or stereotype individuals or communities, perpetuating harmful stereotypes and promoting intolerance (UN, n.d.).

Despite this definition appearing relatively clear and concise, defining hate speech is no easy task. Any attempt at defining hate speech effectively has to try to strike a balance between safeguarding freedom of expression – as a fundamental human right – and protecting individuals and groups from harm. This is further complicated given the fact that different individuals may perceive the same speech act differently, and what one person considers offensive or harmful, another may view as a legitimate expression of opinion. The subjective nature of determining whether specific expressions cross the line into hate speech makes it challenging to establish universally applicable definitions.

Despite these definitional difficulties, it is crucial to engage with the topic, especially in online settings like the metaverse. As we learned from early research into virtual worlds, what happens within these worlds can shape the values that influence individuals' real-world lives, and vice versa. For example, writing in the early 1990s, Christine Ward Gailey argued that video games reflect dominant cultural values in society, reinforcing and promoting behaviors that align with the dominant ideology (Gailey, 1993). In essence, commercially successful games often replicate and reinforce the values and activities associated with prevailing societal norms.

Biases in Virtual Worlds

As a result, it is perhaps unsurprising that the portrayal of different social groups in the virtual worlds of video games tends to reflect existing biases

held by individuals in the real world, including sexism and racism. For example, the gender stereotyping of video game characters' appearances, where women are often depicted as thin with large breasts and emphasized sexual features (a trend epitomized by the character Lara Croft in the Tomb Raider series) and men are often portrayed with muscular physiques and aggressive attitudes, has been well documented. Psychological studies among adolescents and college students indicate that exposure to these stereotypical portrayals can desensitize individuals to real-world sexism, making instances of sexism seem less shocking, potentially perpetuating harmful beliefs such as "rape myths" that blame victims of sexual assault (Breuer et al., 2015).

Of course, sexism in video games is not only an issue when it comes to content. User behavior, particularly within multiplayer online games, is also often deeply problematic. For example, female players are often targeted and harassed and have to develop their own coping mechanisms and strategies in order to safeguard themselves against undesirable behavior.

Racist imagery has also been a persistent issue in video games since the early 1990s. Games like Duke Nukem 3D and Shadow Warrior exemplify how troubling racial stereotypes have shaped the narratives of video games. Duke Nukem 3D's storyline revolves around eugenic panic concerning race mixing between invading aliens and white women in a future Los Angeles depicted as mono-ethnic. The main character embarks on a mission to stop the alien invaders in order to preserve the genetic purity of the human species. In Shadow Warrior meanwhile, the protagonist's stereotypical and generic "Asian" identity is accompanied by a skill set portrayed as biologically determined, perpetuating ideas about the character's perceived deficient masculinity (Weise, 2021).

Sexist and Racist Ideas Have Real World Consequences

The sexist and racist ideas and representations depicted in video games can transcend the virtual world and have real-world consequences. Psychological research suggests that playing violent video games can increase ethnocentrism and trigger heightened aggression when individuals encounter someone who is different from themselves (Ewoldsen et al., 2012). Despite activists like Anita Sarkeesian taking up these issues within the gaming

industry,[1] hate and discriminatory attitudes within video games have become normalized over time, which is likely to influence the design and development of other virtual worlds too. In 2022, media outlets were quick to notice the metaverse's "groping problem", which only goes to show how important it is that we address and challenge these issues in order to create inclusive and safe online environments.

Hate Speech on Social VR

On social VR platforms, such as VRChat, AltSpace, and Meta Horizons, hate speech often takes similar forms to that of more traditional social media. Users can engage in discriminatory or offensive behavior targeting individuals or groups based on their characteristics or identities.

- Racist or derogatory remarks: Users may verbally express racial slurs, engage in racial stereotyping, or make discriminatory comments based on a person's race or ethnicity.
- Homophobic or transphobic behavior: Hate speech can manifest as verbal harassment, bullying, or exclusion, targeting individuals based on their sexual orientation or gender identity.
- Religious or cultural intolerance: Users may engage in hate speech by expressing discriminatory attitudes or insulting remarks against specific religions or cultural groups.
- Cyberbullying and harassment: Like traditional social media, social VR platforms can become spaces for targeted harassment, where individuals are subjected to online bullying, threats, or offensive behavior.

While there are important lessons to be learned from how these forms of behavior take place on social media, there are a number of important differences between them and social VR platforms, particularly given their immersive nature.

Social VR platforms introduce additional elements to user experiences, which can affect the nature and extent of hate speech, bullying, and discrimination. These platforms offer users the opportunity to embody avatars and engage in more immersive interactions. As a result, new or different forms of behavior may emerge. Some examples of this are:

1 Feminist Frequency (2013). Damsel in Distress: Part 1 – Tropes vs Women in Video Games. Online: https://youtu.be/X6p5AZp7r_Q (Accessed on June 27th 2023).

- Nonverbal expressions: Hate speech can extend beyond verbal communication. Users can utilize avatars to engage in offensive or discriminatory gestures, actions, or visual representations, which can amplify the impact of hate speech.
- Spatial proximity and presence: In VR environments, users can physically navigate and interact with others in close proximity. This physical presence can intensify the emotional impact of hate speech, leading to increased feelings of harassment or discrimination.
- Immersive experiences and anonymity: The immersive nature of VR can provide a heightened sense of anonymity and disinhibition, potentially leading to more extreme or offensive behavior compared to traditional social media platforms.

Mitigating Hate Speech on Social VR

To mitigate the negative impact of hate speech in social VR platforms, it is important for platform providers to prioritize proactive moderation, establish clear community standards, and promote user empowerment through reporting tools and educational initiatives. Collaborative efforts involving platform developers, users, and relevant stakeholders can help create safe and welcoming virtual environments where users can freely express themselves without fear of harassment or discrimination.

Recent literature on technology ethics emphasizes the value of deliberative engagement for shaping technological development.[2] These works widely – both implicitly and explicitly – draw on concepts of "deep democracy" (Buhmann and Fieseler, 2021, p. 101475) by highlighting the epistemic potential of open engagement processes. Broadly speaking, this literature proposes that innovators, as proactive participants of a wider public debate and discourse, can contribute to responsible processes of innovation. So essentially, the goal is to harness the potential of different forms of engagement in order to help find optimal solutions.

Around 2010, the term "Responsible Innovation" (RI) became a popular way of talking about responsibility-related issues for academics and policy-makers alike. It is used to refer to a way of organizing research and innovation so that its impacts are safe, equitable, and aligned with societal needs.

2 See, for example, several articles published in the Journal of Responsible Innovation and the Journal of Responsible Technology.

New and emerging technologies, like AI and VR, are going to shape our future in powerful new ways. As we can already see, the result of this is that we will need to confront new questions about risks, ethics, justice, and equity. Responsible innovation essentially provides us with the concepts and practices that are required to address these sorts of questions, helping us to think about things like hype, scale, power, and inclusion in research and innovation.

Drawing on the work of Buhmann and Fieseler, responsible innovation can be seen as encompassing three main dimensions (which are often reflected by the public discourse surrounding new and emerging technologies). First, the *responsibility* to *avoid harm*, which refers, for example, to risk management approaches supposed to control for potentially harmful consequences. Second, the *responsibility to do good,* which refers to the improvement of living conditions, such as are set out in the sustainable development goals. Finally, *governance responsibility,* which refers to the responsibility to create and support global governance structures that can facilitate the former two responsibilities (Ibid).

Examples of the sorts of tools and concepts that fall under the umbrella of responsible innovation are:

- Value Sensitive Design: A framework for exploring stakeholder's values in order to then translate those values into operational design criteria, through iterative conceptual, empirical, and technical investigations. VSD asks questions such as: What values to include in design? How to make these values bear on the design process? How to make choices and trade-offs between conflicting values? How to verify whether the designed system embodies the intended values?
- Scenario Planning Workshops: Narratives, or scenarios, are essentially hypothetical sequences of events constructed for focusing attention on causal processes and decision points. In this sense, the development of scenarios can be used for learning and deliberation, producing decision-making processes that are based on the involvement and interaction of different stakeholders.
- Envisioning Cards: Combines both VSD and scenario planning, the Envisioning Cards are built upon a set of envisioning criteria that are intended to raise awareness of long-term and systemic issues in design. Cards provide prompts for thinking through various implications and value tensions and can be used within workshops or team meetings to trigger discussion and reflection.

It is ultimately in and through these sorts of approaches to technology development that we can try to confront and mitigate potential harms, such as hate speech and bias, before they become locked in.

Discussion

With regards to social VR applications, we can already see the emergence of several key issues that need to be addressed sooner rather than later. For example, managing user behavior in online platforms presents a serious challenge, due to the fragmented domain area and multiple systems involved. Each platform may have its own rules and guidelines, leading to conflicting values and standards. Furthermore, enforcing age restrictions becomes difficult, as users can misrepresent their age or simply bypass the restriction. Another issue is that the speed of development and deployment in the digital space often outpaces the ability to implement effective moderation measures. It is also difficult to legally mandate industry-wide standards, making it hard to establish consistent guidelines for content regulation. Testing these platforms under real-world conditions is also complicated, as the virtual environment is still relatively uncharted territory. In light of these factors, we must ensure that virtual worlds are developed responsibly, which will require continuous adaptation and collaboration between platform developers, users, and regulatory bodies.

Until now, efforts to combat hate speech have largely involved fostering dialogue, promoting education, and encouraging media literacy to enhance understanding, empathy, and respect among individuals and communities. In the case of the development of the metaverse, collaborations between governments, civil society organizations, and technology companies will also play a crucial role in developing guidelines, policies, and tools to address hate speech effectively while respecting diverse perspectives and cultural sensitivities.

It is important to acknowledge that while social VR platforms undoubtedly offer opportunities for socialization and creativity, they also face significant challenges ahead when it comes to addressing hate speech, bullying, and discrimination. As discussed, moderating content and enforcing policies are both complex tasks due to the dynamic and immersive nature of VR environments. However, drawing upon ideas and concepts from responsible innovation, implementing reporting mechanisms, educating

users, and developing community guidelines all can and more importantly *should* play a role in fostering inclusive and respectful virtual communities.

References

Breuer, J., Kowert, R., Festl, R., & Quandt, T. (2015). Sexist games= sexist gamers? A longitudinal study on the relationship between video game use and sexist attitudes. *Cyberpsychology, Behavior, and Social Networking,* 18(4), 197–202.

Buhmann, A. & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society,* 64, 101475.

Ewoldsen, D. R., Eno, C. A., Okdie, B. M., Velez, J. A., Guadagno, R. E., & DeCoster, J. (2012). Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior. *Cyberpsychology, Behavior, and Social Networking,* 15(5), 277–280.

Gailey, C. W. (1993). Mediated messages: Gender, class, and cosmos in home video games. *Journal of Popular Culture,* 27(1), 81.

UN (n.d.). *What is hate speech?* https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Weise, M. (2021*). The Hidden, Destructive Legacy of 'Duke Nukem'.* https://www.vice.com/en/article/pkdvxb/the-hidden-destructive-legacy-of-duke-nukem