Understanding and Mitigating New Harms in Immersive and Embodied Virtual Spaces: A Speculative Dystopian Design Fiction Approach

Guo Freeman
guof@clemson.edu
Clemson University
USA

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

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

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

Lingyuan Li
lingyuanli9936@gmail.com
The University of Texas at Austin
USA

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

Cecilia Aragon*
aragon@uw.edu
University of Washington
USA

Syed Ali Asif*
asifrabi@udel.edu
University of Delaware
USA

Jakki Bailey*
j.bailey@ischool.utexas.edu
The University of Texas at Austin
USA

Meryem Barkallah*
meryemba@umich.edu
University of Michigan-Flint
USA

Braeden Burger*
beburger@umich.edu
University of Michigan-Flint
USA

Sebastian Cmentowski*
sebastian.cmentowski@uwaterloo.ca
University of Waterloo
Canada

Jamie Hancock*
jhancock@turing.ac.uk
The Alan Turing Institute
United Kingdom

Leanne Hides*
l.hides@uq.edu.au
The University of Queensland
Australia

Hongxin Hu*
hongxinh@buffalo.edu
University at Buffalo
USA

Yang Hu*
yhu3@clemson.edu
Clemson University
USA

Wangfan Li*
wangfal@g.clemson.edu
Clemson University
USA

Ruchi Panchanadikar*
rapanch@clemson.edu
Clemson University
USA

Niloofar Sayadi*
nsayadi2@nd.edu
University of Notre Dame
USA

Devin Tebbe*
devinteb@umich.edu
University of Michigan-Flint
USA

Leslie Wöhler*
woehler@hal.t.u-tokyo.ac.jp
The University of Tokyo
Japan

Xinyue You*
xinyueyou23@utexas.edu
The University of Texas at Austin
USA

Zinan Zhang*
zzinan@psu.edu
The Pennsylvania State University
USA

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

ABSTRACT

Seeking novel approaches to understand and mitigate emerging and understudied new harms in immersive and embodied social spaces is a critically needed HCI and CSCW research agenda for achieving safer online environments in the future. In this work, we present and discuss the results from a CHI 2024 workshop in which 26 experts engaged in a speculative dystopian design fiction activity. Through the structured design fiction exercise, our participants

*Equal contributions

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11.
https://doi.org/10.1145/3678884.3681865

collectively created six design fictions along four main themes highlighting our shared concerns in this problem space. We contribute to CSCW and HCI research by demonstrating the novelty and value of using speculation and design fiction as a methodological tool to engage with research in this emerging problem space. The identified themes and created design fictions also help us speculate plausible and desirable futures regarding new harms in embodied and immersive virtual spaces in the first place, which will inform our future research on envisioning and identifying potential solutions to prevent these harms from becoming reality.

CCS CONCEPTS

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

KEYWORDS

toxicity; online harm; online safety; harassment; embodiment; harm mitigation; AI; immersive virtual worlds

ACM Reference Format:

Guo Freeman, Julian Frommel, Regan L. Mandryk, Jan Gugenheimer, Lingyuan Li, Daniel Johnson, Cecilia Aragon, Syed Ali Asif, Jakki Bailey, Meryem Barkallah, Braeden Burger, Sebastian Cmentowski, Jamie Hancock, Leanne Hides, Hongxin Hu, Yang Hu, Wangfan Li, Ruchi Panchanadikar, Niloofar Sayadi, Devin Tebbe, Leslie Wöhler, Xinyue You, Zinan Zhang, and Douglas Zytko. 2024. Understanding and Mitigating New Harms in Immersive and Embodied Virtual Spaces: A Speculative Dystopian Design Fiction Approach. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3678884.3681865

1 INTRODUCTION

Content Warning: In this submission, we explore 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 [8, 31, 64]. Moving towards gaming and virtual world contexts paints an even more dire 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 [19]. Now toxicity is largely accepted and normalized in online gaming [7, 30, 67] with statistics suggesting that 83% of adult gamers experience harassment in online multiplayer games with severe effects such as 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, ranging from *embodied harassment* in social Virtual Reality (VR) [29] to new forms of online racism such as *racist Zoombombing* [45] and *Ugandan Knuckles* in VRChat [2], new AI-powered online attacks such as *hate raids* on Twitch [16], user- or developer-generated harmful design to manipulate people's actions [39], the lack of comprehensive accessibility in XR environments that further marginalizes and harms online users with various disabilities [17, 28, 69], and new types of perceptual manipulations leveraging the core abilities of XR technology to create illusions to the user [14, 33, 63].

We thus 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 and CSCW, including but not limited to: 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. Therefore, in this work, we report our results when leveraging a speculative dystopian design fiction approach to explore this topic with 26 experts at a CHI 2024 Workshop on emerging online harms in immersive and embodied virtual spaces [27]. Through a structured design fiction exercise [26], our participants collectively created six design fictions that highlight four main themes regarding our shared concerns in this problem space, including *monetizing embodied harms, blurring reality with the online world, platforming perpetrators by investigating their motivations and emotions, and embodied harms specifically targeting children.* We contribute to CSCW and HCI research by demonstrating the novelty and value of using speculation and design fiction as a methodological tool to engage with research in this emerging problem space. The identified themes and created design fictions also help us speculate plausible and desirable futures regarding new harms in embodied and immersive virtual spaces in the first place, which will inform our future research on envisioning and identifying potential solutions to prevent these harms from becoming reality.

2 RELATED WORK

In the following, we outline some specific examples of these new potential harms that have been highlighted in existing CSCW and HCI literature.

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) [28, 43]. 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 [57]) and immersive (i.e., being enveloped by, included in, and interacting with the virtual environment [68]) 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 [9, 10, 24, 25, 29, 47, 51–55, 58, 59, 70]. 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 [29]. 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 [57].

New Forms of Online Racism. As HCI continues to advocate an anti-racism/racist research agenda [1, 21], 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" [45]. 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 [32], occurrences of online harassment of Asian Americans during COVID-19 [62], and incidents of racist Zoombombing (i.e., using Zoom, the videoconferencing software, to attack unsuspecting users with racist content [45]). As racism remains a severe and pervasive issue, 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 [61].

New AI-powered Online Attacks. Advanced Artificial Intelligence (AI) technologies, such as AI-based moderation [35, 44, 65, 66], 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 [12, 18, 20, 36, 41, 49]) or by using Natural Language Processing techniques to automatically detect cyberbullying content [6, 50]. 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 [16, 34]. 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 [16, 34]. 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 [39]. 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 [39]. 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 [33]. Tseng et al. [63] and Bonnail et al. [14] demonstrated how developers or content creates might leverage known perceptual manipulation techniques (e.g., redirected walking [48]) to negatively impact the memory or even physical actions of a VR user.

While this small body of existing work has begun to attend to these new harms emerging in various immersive and embodied online spaces, we believe that seeking novel approaches to understand and mitigate these emerging and understudied 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 and CSCW research agenda for achieving a safer online environment in the future. This thus motivates us to leverage a speculative dystopian design fiction approach to explore this topic.

3 METHODS: A SPECULATIVE DYSTOPIAN DESIGN FICTION APPROACH

Speculative thinking has long been used in HCI design and research to anticipate the uncertainty and associated ethical issues of future technologies, uses, and consequences [15], such as through speculative design and design fiction (DF) [4, 11, 13, 23, 40, 42]. The former is a design research framing – a means of speculating about how things could be in possible futures, which supports the society's ability to choose among preferable alternative futures [23]. The latter refers to a specific speculative practice in design research, which not only emphasizes speculation and futurity but also features the distinctive characteristic of fictionality by situating the design in a diegetic world [11, 13, 22, 40, 60]. In this sense, design fiction is "the deliberate use of diegetic prototypes to suspend disbelief about change" [56], which highlights the means of speculating about new ideas through diegetic prototyping in a narrative way while referring to reality [38, 46]. With these understandings, we believe that using such a speculative dystopian design fiction approach is crucial to forming a proactive and pioneering pathway to investigate 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.

In doing so, we conducted a structured design fiction exercise named "Black Mirror Writers' Room" [26] at a CHI 2024 Workshop on emerging online harms in immersive and embodied virtual spaces [27]. This exercise has been successfully used for teaching

and discussing technology ethics through speculation because Black Mirror TV episodes often build upon "our existing anxieties about technology and pushes them just a step farther," which is ideal for "using creative speculation to think through the possible consequences of new technology" [26]. In particular, the core activity of this exercise emphasizes scenario building, which leverages the cognitive diversity of the participants to engage in scenario-writing to envision various social, ethical, and legal implications of future technologies [5, 37].

In this exercise, 26 experts on this topic (14 women and 12 men with diverse ethnicities), including computer scientists, HCI scholars, VR and games researchers, social scientists, and policy makers, first brainstormed collectively on the biggest questions and concerns facing a future in which we interact regularly in immersive and embodied social spaces. They identified potential new online harms in these spaces and then worked in smaller groups to create dystopian design fictions as sample Black Mirror TV episodes using a template. They were instructed to ask themselves: *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?*. After creating their black mirror episode, each group shared their creations back with the entire workshop to discuss the speculative design fictions.

4 RESULTS

Collectively, participants identified four main themes regarding understanding and mitigating new harms in future immersive and embodied virtual spaces based on their shared expertise and interest, including: *monetizing embodied harms, blurring reality with the online world, platforming perpetrators by investigating their motivations and emotions, and embodied harms specifically targeting children.* It should be noted that these themes are not exhaustive and other potential harms were identified (e.g., issues around policy, accessibility, and AI moderation). We do not claim that the issues discussed were the most pressing or important, but they were of the greatest interest to the workshop participants. Built upon these themes, participants created six design fictions using the "Black Mirror Writers' Room" exercise [26], each of which highlighted one or more of the four identified themes.

Theme 1: Monetizing embodied harms. Focusing on this theme, participants created two design fictions. One is titled *"Presidential Strength Challenge"* (see Figure 1). In this story, participants envisioned a future where the new government of the United States has made it illegal to regulate any online activity, claiming it interferes with free speech. As a result, many companies have sprung up to monetize interactions that glorify strength and denigrate any form of sensitivity as weak. The "Presidential Strength Challenge" funds schools that run contests where the winners ("Strong Youth") compete on social media to humiliate their classmates. The more the victims scream and cry the more votes the winners receive. People who complain are reported to the government and ostracized. This story clearly depicts a concerning future where intensified harm in embodied online spaces can be monetized and even normalized to further damage our social and emotional lives. However, participants also went further to ask *"But what happens when the bully and the bullied turn out to be in the same family?"* by highlighting two brothers as the bully and the bullied who did not know each other's true identity in this story. In doing so, participants hoped to emphasize the severe damage of monetizing embodied harms not only on online strangers but on family and close interpersonal relationships.

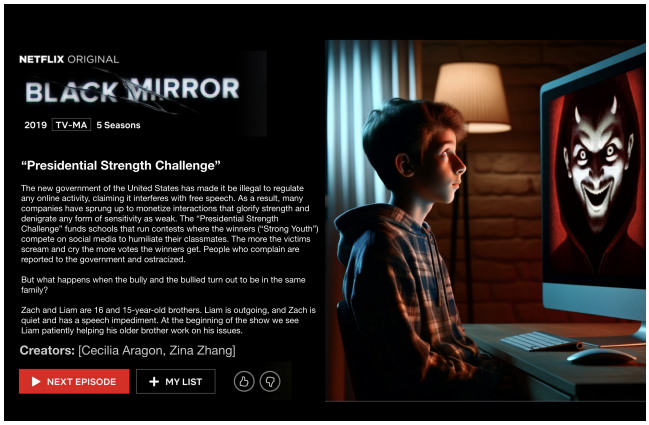

Figure 1: Theme 1 - "Presidential Strength Challenge" Black Mirror episode

The other is titled *"Hijack"* (see Figure 2). In this story, participants wrote about how a young woman named F embeds VR-AR technology directly into her brain, which makes her a renowned avatar in the virtual realm known for her charisma, wit, and beauty. However, her bitter ex, K, hacks into her avatar's control system. Unknown to F, she becomes a mere passenger in her own body as K manipulates her avatar to perform unspeakable acts and utter words she'd never dream of saying. Every humiliating action is also broadcast to the world and monetized. Reflecting upon this story, participants then asked, *"In a world where our avatars are extensions of ourselves, who holds the power to define our true selves?"* Similar to the previous DF, this story highlights how harm in embodied online spaces can be monetized to escalate its damage on the victims. Further, this story also points out the potential risk of further blurring our reality and offline identity with the online world (e.g., hacking one's VR identity could severely damage one's offline identity), an issue further explicated in the next theme.

Theme 2: Blurring reality with the online world. Focusing on this theme, participants also created two design fictions. One is titled *"My Husband's Wife"* (see Figure 3), which tackles new harm in embodied online spaces on intimate, personal relationships. In this story, after moving to a new town Sarah's life is turned upside down by a stalker, Michael, who claims to be her husband. Her life continues to blur into unreality after Michael is arrested, but quickly released after he provides evidence of their "marriage" such as a shared bank account and numerous pictures and videos of their relationship from the official virtual reality platform of their country. They both eventually realize they are at the heart of a complex, long-term scam in which Michael was seduced by and eventually married an AI-generated persona of Sarah built on stolen photos and videos from her social media accounts. Reclaiming their personal and financial independence from the scammers puts them in a battle neither is ready for - because no one believes them. Similar to *"Hijack,"* this DF envisions the continuous collision of our

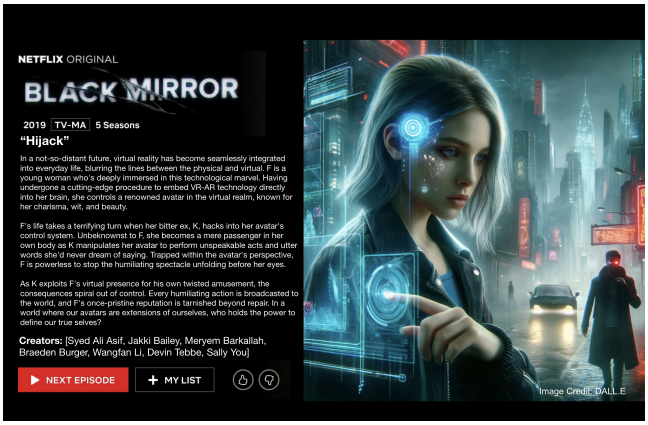

Figure 2: Theme 1 - "Hijack" Black Mirror episode

offline identity and digital representations and asks: can our digital presence fully represent and equal our offline identity? When our digital presence increasingly embodies our offline presence, how can we prove which is real (or more real)?

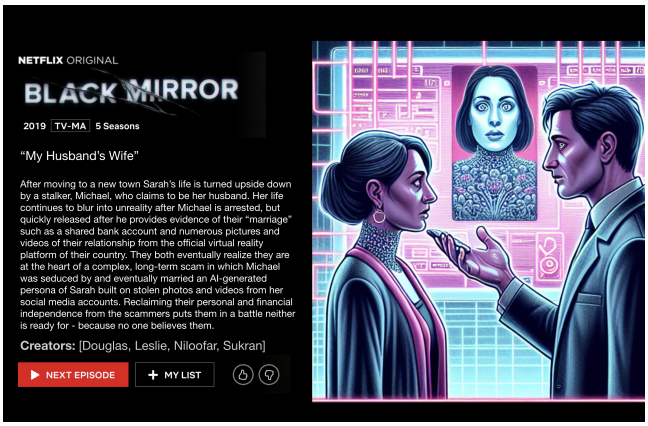

Figure 3: Theme 2 - "My Husband's Wife" Black Mirror episode

The other DF in this theme is titled *"Plug in"* (see Figure 4). This story centers around Rina, a supermarket worker who has a boring offline life and keeps asking "what if my world could be more?" Rina thus depends on Gateway, the ultimate augmented reality platform with a neuro-enhanced user-experience, as a source of entertainment and a way to see their friends wherever and whenever. But when Rina adds an AI-powered plugin to their Gateway that responds to their mood by embedding exciting news and content in their environment, events begin to spiral. Compared to other DFs that emphasize the harm of blurring embodied online identity and offline identity, this DF portrays a disturbing future where an ultimate augmented reality may *overcome* reality (e.g., *"There is no offline, no online."*) and encourages us to reflect upon the prices humans may pay for such a reality (e.g., *"What will it take in return?"*).

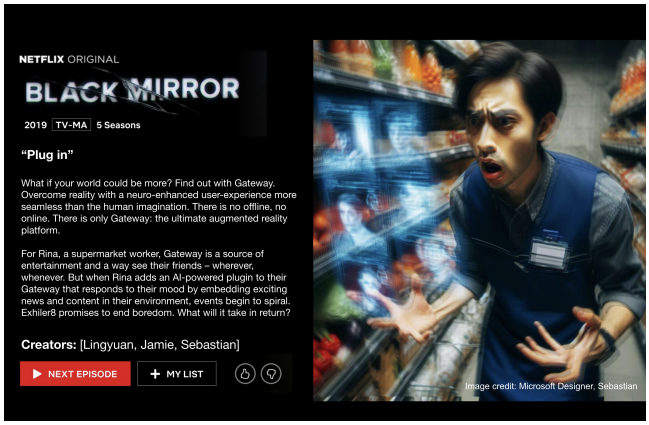

Figure 4: Theme 2 - "Plug in" Black Mirror episode

Theme 3: Platforming perpetrators by investigating their motivations and emotions. Focusing on this theme, participants created a design fiction titled *"Alexithymia"* (see Figure 5). In this story, users are giving social credit for good behavior and intervening in toxicity and bullying in future social VR environments. To rise in the ranks fast, a person named Alex decides to join the virtual world with a second "avatar" which is committing bad behavior and using their primary avatar to intervene. However, it becomes unclear which role Alex is enjoying more. At some point, Alex is acting so viciously towards another avatar that he almost "breaks" the other person trying to convince them to kill themselves. However, participants did not conclude here but added a twist: As Alex is leaving the virtual environment, they suddenly realize that the avatar they left is not the bully but actually the person being bullied. Revealing Alex did not only join the virtual environment two times but every person in this virtual world is a version of Alex who commits all the good and the bad towards themselves. By revealing this twist, participants not only highlight the dystopian and ironic tone of this DF but also reflect upon why toxicity is always part of the online experience. Through this story, participants ask why no matter how we aim to avoid online toxicity, it always comes back.

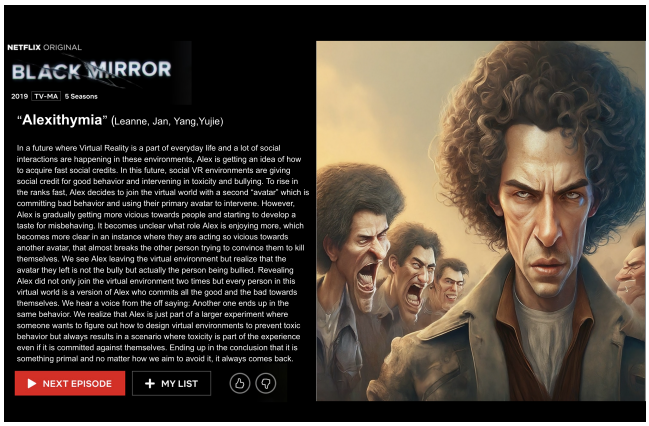

Figure 5: Theme 3 - "Alexithymia" Black Mirror episode

Theme 4: Embodied harms specifically targeting children. Focusing on this theme, participants created a design fiction titled *"School Sucks"* (see Figure 6). This story describes a world where physical schools have been shut down due to teacher shortages and budget cuts. Therefore, children have increasingly learned to work, learn and play in fully immersive online environments and are taught in XR by AI "teachers." These AI "teachers" have "learnt" to force children to aggressively compete in the classroom and treat one another as enemy combatants. Ralph, a 12 year old boy is increasingly dissatisfied with his online interactions and begins to skip class. During one of his truant days he finds a disused youth camp where a group of children who have also exited the online world in favour of being taught by a human ex school teacher. The school teacher, Minerva, is teaching the children to interact with kindness and empathy. However, when a worldwide internet outage occurs due to a climate related disaster, all of the children are forced back into the non-digital world. This results in an apocalypse in which all the violence and toxicity of the online world is suddenly transferred to the real world with children fighting over resources and attacking one another. The camp is found, which devolves into a *hunger-games* style bloodbath. Overall, this story features a futuristic context especially crucial for children (e.g., schools and teachers) and asks how embodied harm in this context can affect and damage children: would children be able to learn appropriate behavior and social norms in virtual spaces and through AI teachers? How would placing children in such technical environments affect social dynamics and our future generation? Can children understand the harm of directly translating their online behaviors in embodied virtual spaces to the offline world?

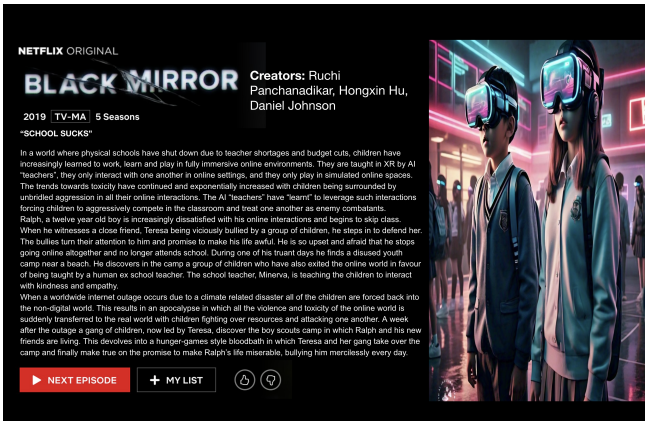

Figure 6: Theme 4 - "School Sucks" Black Mirror episode

5 DISCUSSION AND FUTURE WORK

Realizing the urgent need to understand and mitigate new harms in immersive and embodied virtual spaces, in this work, we have reported results of our speculative dystopian design fiction approach to explore this topic with 26 experts with diverse research backgrounds. Through a structured design fiction exercise [26], our participants collectively created six design fictions that highlight four main themes regarding our shared concerns in this problem space.

Indeed, the CSCW, HCI, and design research communities have all in their own ways recognized the potential of speculation for contributing towards responsible technology design and meaningful social change in the future. Our results further show that this speculative dystopian design fiction approach provides a valuable methodological tool to engage with our ongoing research efforts on understanding emerging harm in future embodied and immersive virtual spaces and identifying potential solutions to mitigate them.

We do so in several ways. First, our participants created **narrative** DFs, which combines "the traditions of writing and story telling with the material crafting of objects" [11]. In this sense, these stories are more than a literary genre but express a design practice in terms of its rules, contexts, boundaries, and affordances (e.g., what can go wrong in embodied and immersive virtual spaces in a not-so-distant future?). Second, our participants' DFs engage with **the social world**. DFs create "socialized objects" that tell stories about "new, different, distinctive social practices that assemble around and through" them [11]. As shown in our results, all stories were built upon our shared knowledge and concerns about existing issues in the current embodied and immersive virtual spaces. However, they also go beyond to highlight new and even more problematic behaviors in the future if no appropriate action is taken. Therefore, these created DFs are also **provocative and encourage reflection**. All 6 created DFs are not attempting to resolve existing issues and concerns but to provoke "what if" questions that reflect on the deeper social, cultural, political and ethical implications of technologies in a long term [23, 40]: What if embodied harm can be monetized? What if our reality and the online world continue to blur (do we even still have a "reality" then?)? What if we never get to understand perpetrators' motivations and emotions (can we really get rid of online toxicity without understanding these?)? And what if placing children in these technical environments irrevocably damages our future generation?

Although our focus is not to immediately address these questions, our speculative dystopian design fiction approach helps to speculate plausible and desirable futures regarding new harms in embodied and immersive virtual spaces in the first place, which will also help to envision and start identifying potential solutions to prevent these harms from actually happening. Therefore, for future work, we plan to (1) continue to identify other potential harms (e.g., issues around policy, accessibility, and AI) in future immersive and embodied virtual spaces that are not extensively covered in this work; and (2) convene another workshop with experts in this problem space to focus on envisioning potential and plausible solutions based on these identified themes and design fictions. We hope that leveraging this speculative dystopian design fiction approach will contribute to the evolving research focus on understanding and mitigating new harms in future embodied and immersive virtual spaces in a proactive way.

ACKNOWLEDGMENTS

We thank CHI 2024 for hosting our workshop [27] and for making this work possible. We especially thank Yujie Tao and Sukran Karaosmanoglu for contributing to the workshop activities. This work was partially supported by the National Science Foundation awards 2112878 and 2211896.

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] Jeffrey Bardzell and Shaowen Bardzell. 2014. "A great and troubling beauty": cognitive speculation and ubiquitous computing. *Personal and ubiquitous computing* 18 (2014), 779–794.

[5] Julia Barnett and Nicholas Diakopoulos. 2022. Crowdsourcing impacts: exploring the utility of crowds for anticipating societal impacts of algorithmic decision making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 56–67.

[6] 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.

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] Julian Bleecker. 2022. Design fiction: A short essay on design, science, fact, and fiction. *Machine learning and the city: Applications in architecture and urban design* (2022), 561–578.

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

[13] Mark Blythe. 2014. The hitchhiker's guide to ubicomp: using techniques from literary and critical theory to reframe scientific agendas. *Personal and ubiquitous computing* 18 (2014), 795–808.

[14] 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

[15] Philip Brey. 2017. Ethics of emerging technology. *The ethics of technology: Methods and approaches* (2017), 175–191.

[16] 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).

[17] Chris Creed, Maadh Al-Kalbani, Arthur Theil, Sayan Sarcar, and Ian Williams. 2024. Inclusive AR/VR: accessibility barriers for immersive technologies. *Universal Access in the Information Society* 23, 1 (2024), 59–73.

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] Paul Dourish and Genevieve Bell. 2014. "Resistance is futile": reading science fiction alongside ubiquitous computing. *Personal and ubiquitous computing* 18 (2014), 769–778.

[23] Anthony Dunne and Fiona Raby. 2024. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming*. MIT press.

[24] 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.

[25] 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.

[26] 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.

[27] 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*. 1–7.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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

[34] 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.

[35] 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)*.

[36] 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.

[37] Kimon Kieslich, Nicholas Diakopoulos, and Natali Helberger. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment. *AI and Ethics* (2024), 1–23.

[38] David Kirby. 2010. The future is now: Diegetic prototypes and the role of popular films in generating real-world technological development. *Social Studies of Science* 40, 1 (2010), 41–70.

[39] 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

[40] Conor Linehan, Ben J Kirman, Stuart Reeves, Mark A Blythe, Theresa Jean Tanenbaum, Audrey Desjardins, and Ron Wakkary. 2014. Alternate endings: using fiction to explore design futures. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 45–48.

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

[42] Jonathan Lukens and Carl DiSalvo. 2012. Speculative design and technological fluency. *International Journal of Learning and Media* 3, 4 (2012).

[43] 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*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300794

[44] 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.

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

[46] Amrita Narlikar. 2007. All that glitters is not gold: India's rise to power. *Third World Quarterly* 28, 5 (2007), 983–996.

[47] 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

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

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

[50] 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.

[51] 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.

[52] 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* 7, CSCW2 (2023), 1–29.

[53] 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 Conference on Human Factors in Computing Systems*. 1–17.

[54] Kelsea Schulenberg, Lingyuan Li, Caitlin Lancaster, Douglas 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* 7, CSCW2 (2023), 1–30.

[55] 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.

[56] Slate. 2012. *Sci-Fi Writer Bruce Sterling Explains the Intriguing New Concept of Design Fiction*. https://slate.com/technology/2012/03/bruce-sterling-on-design-fictions.html

[57] 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.

[58] 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

[59] 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/

[60] Theresa Jean Tanenbaum. 2014. Design fictional interactions: why HCI should care about stories. *interactions* 21, 5 (2014), 22–23.

[61] 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.

[62] 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.

[63] 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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] Kexin Zhang, Elmira Deldari, Yaxing Yao, and Yuhang Zhao. 2023. A Diary Study in Social Virtual Reality: Impact of Avatars with Disability Signifiers on the Social Experiences of People with Disabilities. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–17.

[70] 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.