"Comforting and Small Like a House Cat, Big and Intimidating Like a Bodyguard": How Women Perceive and Envision AI Companions as a New Harassment Mitigation Approach in Social VR

Guo Freeman
School of Computing
Clemson University
Clemson, South Carolina, USA
guof@clemson.edu

Kelsea Schulenberg
School of Computing
Clemson University
Clemson, South Carolina, USA
kelseas@g.clemson.edu

Lingyuan Li
School of Information
The University of Texas at Austin
Austin, Texas, USA
lingyuan.li@ischool.utexas.edu

Ruchi Panchanadikar
School of Computing
Clemson University
Clemson, South Carolina, USA
rapanch@clemson.edu

Nathan McNeese
Human-Centered Computing
Clemson University
Clemson, South Carolina, USA
mcneese@clemson.edu

Abstract

Companionship is crucial for people's everyday psychological well-being. With growing concerns over harassment against women in embodied social VR spaces, we turn an eye towards AI companions as a potential new approach to protect women in social VR by better fulfilling their under-addressed harassment mitigation needs. Using 20 interviews with women social VR users, we reveal their envisionings for leveraging AI as *Accessible Companions*, *Informational Companions*, *Emotional Support Companions*, and *Protective Companions* to better protect them in social VR compared to their existing safety mechanisms and strategies. We also reflect upon various sociotechnical complexities for designing and implementing such AI companions in social VR spaces and propose three design principles to inform future efforts to create AI companions to protect women and other marginalized users in social VR. Our work contributes to ongoing discussions on nuanced harassment mitigation approaches that further support marginalized social VR users' multidimensional needs without harming their self-agency, human relationships, and supportive networks.

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Keywords

artificial intelligence, AI companion, online harassment, women, online safety, social VR

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713473

ACM Reference Format:
Guo Freeman, Kelsea Schulenberg, Lingyuan Li, Ruchi Panchanadikar, and Nathan McNeese. 2025. "Comforting and Small Like a House Cat, Big and Intimidating Like a Bodyguard": How Women Perceive and Envision AI Companions as a New Harassment Mitigation Approach in Social VR. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3713473

1 Introduction

The HBO documentary "We Met in Virtual Reality," filmed entirely in the most widely-used social Virtual Reality (VR) platform VRChat, captures the growing power and potential for social VR to enhance people's digital lives in unique ways [31]. Indeed, unlike in traditional online social spaces where a screen mediates one's social interactions, social VR allows users to interact in **embodied** (i.e., experiencing a virtual bodily representation as one's own [78]) and **immersive** (i.e., feeling enveloped by and included in the virtual space [90]) ways. These interactions can be facilitated through a unique combination of features, including the use of VR head-mounted displays, partial or full-body-tracked avatars that mirror real-time movements, synchronous voice conversations, and simulated touching features [25, 53, 71–73].

However, there has also been increased scrutiny on how social VR platforms (e.g., VRChat, Rec Room, Bigscreen, and Meta Horizon Worlds) may facilitate more embodied and immersive harms, which can be felt in more realistic, physicalized, and severe ways compared to harassment on traditional 2D-based online platforms. These harassing behaviors in social VR especially target marginalized users such as women and particularly women of color, which have been frequently documented in mass media and technology reports, such as trash-talking women, drawing penises, virtual "groping" and "rape," and the most recent "gang rape" of a teenage girl's avatar in the metaverse [8, 9, 27, 57, 71, 76, 79, 80, 85]. In recognition of these harms, women social VR users have actively leveraged various platform-specific safety features (e.g., muting, blocking, reporting, and personal space bubbles) [8, 9, 27, 71] and personal

strategies (e.g., hanging out in a group of people, avoiding head-on confrontation, and even "harassing" their harassers back) [71] to protect themselves from harassment in social VR spaces. Yet, they have collectively revealed three major limitations of these existing safety mechanisms and methods for women, including: (1) *reinforcing existing unequal power dynamics in social VR,* a still largely cisnormative and male-dominated social space; (2) *offering limited informational, social, and emotional support to women victims who have been harassed*; and (3) *significantly reducing women's embodied experiences in social VR* [8, 9, 27, 57, 71, 76, 79, 80, 85].

Therefore, while social VR becomes increasingly important for our future networked social lives, it continues to reinforce and amplify harassment risks for marginalized online users such as women. As such, we believe that more research is critically needed to empirically investigate how to better protect women in these spaces via more nuanced harassment mitigation approaches, as women largely view existing methods as insufficient to prevent these intensified forms of harassment against them.

In this paper, we particularly focus on how women social VR users perceive and envision one of the most recent technological advances - **AI as companions** - can be leveraged to help them mitigate harassment risks in social VR spaces, especially compared to existing safety mechanisms and strategies they have used. We chose this focus for two reasons.

First, companionship is crucial to building key human social dynamics of knowing, loving, and caring for a person [36]. In social scientific literature, *companionship* often refers to an inner and intimate social bond between two individuals that provides mutual enjoyment and emotional satisfaction, improves psychological well-being, and reduces life stress and feelings of loneliness [66]. However, it can be extended beyond human relationships to include companion animals (e.g., pets [5]) and even non-living entities (e.g., toys and children's imaginary companions that provide comfort and company [34]). One recent example of such non-living entities is AI companions (e.g., *Replika* [1]), which can be broadly defined as "virtual conversational agents that possess a certain level of intelligence and autonomy as well as social skills, allowing them to establish and maintain long-term relationships with users" using natural language [44, 82]. They have long been used within different gaming and virtual world contexts to fulfill players' social and emotional needs (e.g., non-playable characters [NPCs] in games to protect players and make players feel safer) [20, 28, 52, 91], and continue to (re)shape how we define companionship in the modern digital society [65, 82]. Meanwhile, while existing works have begun to explore how AI technologies can be designed to mitigate new forms of harassment in social VR (e.g., through AI-based moderation systems [72] and AI moderator avatars for protecting children [22–24]), there still exists an urgent need to explore further how AI can be leveraged to go beyond the automation of moderation efforts to provide for under-addressed harassment mitigation needs. Such needs include the need for emotional and esteem support after a harassing incident has occurred, which has been especially highlighted in women's experiences with harassment in social VR [57, 71, 79, 80, 85]. Therefore, we believe that designing new AI companions may help address this critical need to better support women social VR users.

Second, as mentioned above, women largely find existing platform-specific safety features and their own personal strategies to mitigate harassment against women in social VR ineffective. It seems important to explore alternative harassment mitigation approaches that can further enhance women's existing strategies while addressing their limitations, such as allowing women to access help at any time without forcing them to rely on other human user groups and unequal power dynamics in social VR. In this sense, designing new AI companions may become a promising new alternative due to the automatic and computational nature of such technologies, which requires more in-depth investigations.

Drawing upon 20 in-depth interviews with women who use various social VR platforms, we aim to address the following two research questions in this work:

RQ1: How do women envision the potential of leveraging AI companions to protect them from harassment in social VR, especially compared to existing safety mechanisms and strategies they have used?

RQ2: How do women perceive potential challenges with using AI companions as a new harassment mitigation approach in social VR?

We contribute to HCI research at the intersection of AI, online safety, and social VR in three ways. First, we contribute to the growing body of literature on harassment and harm in embodied social VR spaces (e.g., [8, 9, 22–24, 27, 67, 76, 93]) by especially exploring more nuanced harassment mitigation approaches through AI companions to better protect women, who are often considered marginalized in largely cisnormative and male-dominated social VR spaces. This further confirms and expands insights of women's struggles with not only severe harassment risks but also with the lack of sufficient safety mechanisms to protect them from such harassment in social VR. By centering women's own envisions and perceptions, this work gives women a voice in shaping future technologies to protect them. Second, AI companions are not yet provided to users in any form by social VR platforms and are not explicitly crafted to fulfill these harassment mitigation roles in other online contexts. Therefore, our work offers a particularly pioneering approach to exploring novel ways to leverage AI that go beyond a moderator or a punisher to better fulfill marginalized social VR users' under-addressed harassment mitigation needs (e.g., women's need for informational, social, and emotional support when dealing with harassment). Third, our work additionally provides unique insights into various sociotechnical complexities that must be carefully considered when designing and implementing AI companions in social VR spaces. Based on women's own envisions and our critical reflections, we propose three foundational design principles that can inform future efforts to create effective and sensitive AI companions to protect women and other marginalized users in social VR spaces.

[1] https://replika.com/

2 Related Works

2.1 Women's Severe Challenges to Deal with Harassment Online and in Social VR

Although both men and women can be harassed online, women are twice as likely as men in the United States to say that their most recent incident of harassment was very or extremely upsetting, and harassment behaviors targeting women such as physical threats and stalking are on the rise [88]. This appears to be a global issue, as women in South Asia (e.g., India and Pakistan) have reported similar abuses [68, 86, 87]. The gaming and virtual world contexts paint an even more concerning picture, as women are consistently marginalized in these male-dominated and often competitive communities [26, 48, 49, 59]. Some of the earliest documentation of online harassment against women in virtual worlds can even be traced as far back as the early 1990s [18].

Social VR, as the more recent immersive and embodied social technology, further contributes to the evolution of online harassment against women. A growing body of HCI and CSCW research has warned that, while social VR's unique combination of features for embodiment and immersion may lead to harassment risks for all social VR users (e.g., allowing people to virtually "touch" or assault others), individuals who present certain offline identity characteristics (e.g., women, LGBTQ+, and racial minorities) in social VR through avatar design, voice, or nonverbal cues such as body language could face an even greater risk [8, 9, 27, 57, 58, 76]. For example, a 2017 technology report with 13 social VR women users reveals several safety risks for women in social VR, such as sexual harassment and flirting [57]. A 2018 report shows that 49% of women respondents reported having experienced at least one instance of sexual harassment [58]. Freeman et al. highlight that harassment in social VR may be felt as more disruptive to women, LGBTQ+ individuals, and racial minorities because consistent identity practices make it is easier to identify and target them as potential victims [27]. Blackwell et al.'s works also shed light on the risks of sexual harassment and stalking these populations face in social VR [8, 9]. In particular, Schulenberg et al.'s interview study with 31 women social VR users reveals women's significant challenges in managing harassment risks in social VR [71]. They show that while women have developed their own methods to manage these harassment risks in social VR, including both platform-specific tools and creative inter- and intrapersonal strategies, many of these features and strategies introduce limitations and potential harms that may further marginalize women in social VR spaces.

Taken together, these prior works have highlighted three major limitations of existing safety mechanisms and methods for protecting women in social VR, including (1) **Limitation 1: reinforcing existing unequal power dynamics in social VR**: women's personal strategies, such as always hanging out with a group of people or relying on friends who are often not considered marginalized in social VR (e.g., cis straight men) for protection, puts excess burdens on women to build their own safety mechanisms. In doing so, women must find and build friendships with others who they can always access in case they encounter harassment in social VR. These strategies may even inevitably reinforce the already unequal power dynamics in social VR, as they suggest that women must rely on men to protect themselves in those online spaces; (2) **Limitation 2: offering limited informational, social, and emotional support to women victims who have been harassed**: In fact, there is no existing mechanisms to educate women how to identity and manage harassment when engaging in social VR. There is also no effective strategy to help women overcome the internalized shame and fear stemming from being harassed in social VR other than just performing personal resilience (e.g., building a "thick skin"); and (3) **Limitation 3: significantly reducing women's embodied experiences in social VR**: For example, the use of personal space bubbles to prevent physicalized sexual harassment targeting women unfortunately also limits women's opportunities to interact with other social VR users in an immersive, engaging, and embodied way. Crafting avatar design and voice to hide women's own identities makes women less identifiable in social VR but come at a large cost to their abilities to fully embody their identities [8, 9, 27, 57, 58, 71, 76].

These foundational studies on the voices of women social VR users and their critical challenges to deal with harassment in social VR directly motivated our work. Indeed, women are not only dealing with severe harassment risks in social VR but also facing the unfortunate fact that existing methods, either platform-provided tools or women's own strategies, are insufficient to protect them from such harassment. In response, we pay specific attention to seeking more nuanced harassment mitigation approaches beyond existing safety mechanisms and methods to foster safer social VR spaces for women and other marginalized communities alike. The following section, then, explores the potential of leverage AI companions as a new harassment mitigation approach.

2.2 The Potential of Leveraging AI Companions to Better Protect Women in Social VR

The Current Focus on AI-Based Content Moderation as a Main Harassment Mitigation Approach and Its Limitations. AI systems and tools have been widely used to create safe online environments. For instance, AI has been used to automatically filter certain keywords to block posts or comments that include specific harassing terms and phrases, such as the AutoModerator bot on Reddit [10, 17, 19, 37, 45, 63] and flagging systems used in gaming [40, 81]. Moving to social VR spaces, significant efforts have also been made to explore how to leverage advanced AI technologies, especially AI-based content moderation (e.g., the use of machine learning and decision-making to monitor online spaces for violations and incidents of harassment [32, 37]), to help mitigate social VR harassment in more effective and efficient ways compared to traditional harassment mitigation models (e.g., human-based moderation or text detecting). For example, Schulenberg et al. re-envision AI's new roles in innovating content moderation approaches to better combat harassment in social VR by helping make consistent judgments regarding harassment, managing social VR harassment in real-time and at a large scale, and overcoming potential subjective biases of individual human moderators [72]. Fiani et al.'s work investigates the use of an AI-powered moderator avatar to combat harassment against children in social VR [22–24]. Their findings especially show that children felt safer in the AI moderator's presence,

especially when the AI moderator took on a human-like appearance compared to when the avatar was not present or not visible.

However, prior works have also warned that AI-based content moderation as a platform governance approach may suffer from several limitations. Examples include: a lack of understanding of the nuanced sociocultural context surrounding online harassment [56, 72, 84]; the potential for discriminatory, biased, and nontransparent decision-making that further complicates existing issues of fairness and justice in online social spaces [1, 21, 29, 30, 56, 72, 84]; the risk of enacting "prior censorship" (i.e., one must seek approval from some empowered third party before they are allowed to speak or to publish their views) that violates international human rights law [29, 46]; and technical difficulties in keeping up with evolving content and new forms of online harassment [72]. Therefore, previous literature has highlighted the importance of incorporating human oversight in AI-based moderation to mitigate these issues, such as implementing human-AI collaboration through conditional delegation [41], community engagement via direct and indirect feedback processes [72], and mandatory notice-and-comment procedure for the public to understand and assess AI's decision-making [56].

AI Companions as a New Lens to Protect Women in Social VR. Given these limitations of the current focus on AI-based content moderation as a main harassment mitigation approach, there is a clear need for HCI researchers to explore further how AI can be designed in ways that simultaneously build upon and go beyond current envisionings of AI primarily as a moderator for harassment mitigation in social VR. Drawn upon these existing works and foundational studies on women's experiences with harassment in social VR detailed in section 2.1, we are motivated to explore the potential of designing future AI as **companions** to better protect women from social VR harassment based on the following considerations.

First, companionship is especially valuable for people's everyday psychological well-being, and AI companions seem to further expand how people can seek and experience such companionship in modern digital society. At the high level, companionship is described as a strong, long-term social bond between two individuals who are familiar with each other and continue to spend time together to provide mutual enjoyment and emotional satisfaction [14, 66]. With the rapid technological advances in the field of artificial intelligence, humans also seem to extend such intimate relationships and emotional connections to non-living entities like AI companions (i.e., virtual conversational agents that can establish and maintain long-term relationships with human users through natural language to meet their various social and emotional needs [41, 82, 94]). Indeed, AI companions have been used to support a wide range of human activities and experiences as a personalized, friendly presence to foster a sense of safety and comfort. Successful examples of such commercial products include *Replika*, which is labeled as an AI companion who is "always on your side" and "is always ready to chat when you need an empathetic friend"[2]. Several studies have revealed that AI chatbots serve as valuable companions to enhance one's sense of belonging while mitigating loneliness because they can provide human users with substantial affective and social value [15, 60, 77]. Previous works have also begun to explore how to design such human-AI companionship to support human psychological and emotional needs, including principles of human-like design, adaptivity and adaptability, proactivity and reciprocity, among others [65, 82]. With this growing research agenda on understanding and designing AI companions for improving people's mental well-being, we believe that AI companions show potential to help social VR users, especially marginalized individuals such as women, feel more comfortable and safer in virtual environments.

Second, AI companions have long been used in the virtual worlds and gaming context to help online users feel safe and supported. Indeed, one of the primary ways to embed AI companions into virtual environments such as gaming is through the form of NPCs (non-player characters) that accompany players throughout their gameplay journey as sidekicks or allies. This long-standing tradition is rooted in how these companions are used to support and facilitate players' experiences through diverse roles such as combat participation, guidance provision, emotional engagement, and strategic adaptation [20, 28, 52, 91]. Such AI companions are often designed to manifest genuine emotions and mirror human-like responses. Particularly for humanoid AI companions, the ability to express a spectrum of human-like emotions such as joy, frustration, and curiosity through the design of dynamic facial expressions are essential for forming genuine player-companion connections [33, 42, 74]. As such, AI companions as "sidekicks" impart advice and offer overarching guidance, which helps enrich the narrative experience and emotional support for the player. And AI companions as "allies" are invaluable assets during combat and other pivotal scenarios, which helps elevate the player's capabilities, sense of safety and confidence, and chances to win in the game [20]. Due to this well-established tradition to leverage AI companions to support players in virtual contexts, it seems particularly valuable to explore how AI companions can be used to prevent and mitigate the social and emotional damages of harassment against women social VR users through offering these critically needed advice and guidance, emotional support, and protection.

Third, AI companions may provide alternative approaches to further improve existing safety mechanisms and strategies women have used in social VR and reduce their limitations. As described in section 2.1, existing safety mechanisms and strategies to protect women social VR users not only put an excessive burden on women to access support or limit how women can experience social VR at its full capacity, but also fail to provide effective social and emotional support that a woman victim would critically need after being harassed. As AI companions can be accessed at any time and have been widely used to guide, protect, and support human users in various ways in virtual contexts, how, if at all, future AI companions can be designed to supplement and innovate existing methods to better protect women in social VR spaces presents a unique and understudied research opportunity. Additionally, other prior works tend to craft AI companion designs for harassment mitigation based primarily on deductive testing of pre-made designs or on researchers' interpretations of social VR observations (e.g., [22–24, 93]). More research is still needed to re-emphasize actual envisionings and expectations of women social VR users for designing AI companions to further improve their existing safety strategies.

[2]https://replika.com/

3 Methods

Recruitment and Participants. This study was a part of a multi-year research project on social experiences in social VR. The University's Institutional Review Board (IRB) approved this study for research ethics before recruiting participants. Since this broader project focuses on how diverse users experience social VR, especially those who are often considered marginalized in technology spaces, such as women and LGBTQ individuals, we posted recruitment messages on various popular online forums for social VR users and queer gamers (e.g., r/VRchat, r/Recroom, and r/gaymers in Reddit) and social media platforms (e.g., Facebook and Twitter) to recruit participants who engage in various social VR platforms. In our recruitment messages, we invited individuals who were active social VR users (e.g., had used at least one social VR platform for at least one hour in the past 6 months); had experienced, witnessed, or been accused of harassment in social VR; and were 18 years or older to participate.

We then interviewed all individuals who responded to our recruitment messages and were willing to participate in March and April of 2022 (N=39) as part of the broader project via text/voice chat over Discord or video chat over Zoom, depending on participants' modality preferences. All participants self-reported that they had experienced harassment in social VR. For this study, we used the interview data from all participants who self-identified as women or feminine (N=20) out of the 39 participants. The average age of our participants is 27.63 (excluding 1 N/A response). The majority of our participants are users of VRChat (N=15), AltspaceVR (N=5), and Rec Room (N=3). Additionally, less than three participants use Meta Horizon Worlds, Spatial, Immersed, Bigscreen, and Inverse. On average, our participants had been engaging in social VR for three years and four months with variations from 1 hour to 30 hours per week, which represents a sample of frequent and experienced social VR users. Table 1 summarizes participants' demographic information, social VR experiences and frequency of use, and social VR platforms they mainly use.

Interviews. Before the interviews, we provided an informed consent document to all participants based on their communication preferences (e.g., email or Discord message). We did not collect names or identifiable information from participants. Interview questions were crafted using dialogic techniques designed to encourage participants to engage deeply with their responses [89]. These questions, as detailed further below, drew inspiration from prior literature on social VR and harassment in social VR, particularly from the works of Blackwell et al. [7–9] and Freeman et al. [27], as well as from our own prior experiences with social VR as both researchers and users.

Interviews first began with introductions, basic demographic questions, and questions regarding their level of experience in social VR as well as experiences with harassment in social VR to orient the conversation towards the potential role of AI companions in mitigating social VR harassment. As shown in existing works, there is a distinct lack of consensus among users on what should be defined as harassment in social VR. For instance, Blackwell et al. describe harassment in social VR as subjective and extremely personal, ranging from verbal attacks to violations of physical and personal space [8]. In Freeman et al.'s work, social VR users seem to view "any interaction or experience that intentionally upset them and cause harm, aggravation, anxiety, and instability" as harassment [27]. In different contexts, what might feel harassing to one person might not feel harassing to another when it happens to them, such as the mismatch between children's and adults' expectations for appropriate social behaviors in social VR [50, 51]. Therefore, we did not offer a specific definition of harassment to avoid misleading our participants but encouraged them to freely recount and share as much detail as they felt comfortable and appropriate regarding how they personally experienced harassment in social VR (e.g., *"Please explain to me how you define harassment in social VR?"* and *"How did that experience make you feel about yourself? About social VR?"*). Although these questions are not the focus of this research and therefore are not reported in this paper, they served to orient our participants to reflect on AI companion as a potential approach to mitigate harassment in social VR.

Next, participants were asked to think about how they usually deal with harassment in social VR to understand existing strategies and their limitations to better contextualize the idea of AI companion (e.g., *"Please describe for me strategies that you have used to prevent harassment or have seen other people use to prevent harassment."*). Moving towards the heart of this research project, participants were then asked to describe any new strategies for mitigating harassment that they might find to be beneficial. Before we began specific conversations about AI companions, we provided participants with a brief explanation of AI with the following definition given to participants for interview orientation purposes:

"In short, we can define AI as 'the ability of a machine or a computer program to think and learn. The concept of AI is based on the idea of building machines capable of thinking, acting, and learning like humans.' Some common examples of AI would be Siri or Google Assistant, a computer-controlled opponent in games such as an NPC or a 'boss,' or an enemy in League of Legends."

Built upon this brief definition of AI, participants were encouraged to further envision such AI as a personal **companion** that would interact with them through natural language and accompany them throughout their social VR engagement by providing them with comfort and a sense of safety [41, 65, 82]. It is important to note that although some of our participants had occupations or were in schoolwork related to a technology sector (N=3) (e.g., software developer, P3; PhD student whose research includes creating VR games, P15; content developer, P20), most participants did not have any specific experience building or developing AI technology. Given this, we encouraged our participants to envision such AI companions primarily based on their experiences with digital companions or AI companions in other online contexts (e.g., chatbots or NPCs as sidekicks or allies in gaming) rather than technical building experience. As social VR continues to attract diverse users, it is indeed expected that most users will not be AI experts. Therefore, our sample represents how *actual* women social VR users perceive, envision, and approach the future of AI companions as a new harassment mitigation approach to protect them and other marginalized communities alike in social VR.

Grounded in their envisions of AI as companions in social VR, participants were subsequently asked to describe how they perceived the role of such AI companions in preventing and mitigating social VR harassment (e.g., *"How do you feel about your experiences*

ID	Gender	Age	Ethnicity	Sexuality	Location	Social VR Experiences	Social VR Platform Mainly Used	Self-Reported Usage Frequency (per week)
P1	Woman	25	Black	Lesbian	USA	3 years	Rec Room, AltspaceVR	regularly
P2	Woman	25	Black	Lesbian	USA	8 years	VRChat, Spatial	19 - 23 hours
P3	Woman	24	Black	Straight	USA	3 years	VRChat, Spatial	less than 10 hours
P4	Woman	24	Black	Lesbian	USA	6 years	VRChat	3 - 4 times
P5	Trans Woman	26	White	N/A	USA	3 years	AltspaceVR	very often
P6	Trans Woman	27	Biracial White/Indigenous Canadian	Asexual	Canada	6 months	VRChat	20 hours
P7	Woman	25	Black	Lesbian	USA	5 years	VRChat	more than 20 hours
P8	Genderqueer Feminine Presenting	N/A	Biracial White and Black	N/A	USA	3 years	VRChat	2 hours
P9	Woman	26	Black	Queer	USA	4 years	Rec Room, AltspaceVR	20 hours
P10	Woman	20	Biracial Black and Italian	Asexual	USA	1 year	VRChat	14 - 20 hours
P11	Woman	25	White	Pansexual	Germany	4 years	VRChat	20 hours
P12	Woman	24	Black	Bisexual	N/A	3 years	VRChat	24 - 30 hours
P13	Woman	27	Black	Lesbian	USA	3 years	VRChat	more than 10 hours
P14	Woman	29	Black	Lesbian	USA	5 years	Rec Room	more than 12 hours
P15	Woman	31	Middle Eastern	Lesbian	USA	1 year	VRChat	3 - 4 times
P16	Woman	44	White	Straight	USA	5 years	AltspaceVR, Immersed, Bigscreen, Meta Horizon Worlds	It depends
P17	Woman	40	Biracial Native and Hispanic	Lesbian	USA	4 months	VRChat, Meta Horizon Worlds, AltspaceVR, Inverse	1 - 5 hours
P18	Woman	28	Native and White	Straight	USA	3 years	VRChat	4 - 5 hours
P19	Woman	30	White	Asexual	USA	6 months	VRChat	2 - 5 hours
P20	Woman	25	Black	Straight	USA	2 years	VRChat	a minimum of 14 hours

Table 1: Demographic information and social VR experiences of participants. Note: N/A – participant did not provide information.

in social VR when accompanied by an AI companion?") and to reflect on AI companions in various dimensions. This included a comparison between AI and human companions regarding their efficacy in mitigating harassment (*"How do you think the effectiveness of using an AI companion in mitigating potential harassment in social VR compares to that of a human companion? Why?"*), and expectations of AI companions' behaviors (*"What do you want your AI companion to do [what types of activities?] to mitigate harassment?"*). Finally, participants were asked to envision how they would design AI companions to prevent emergent harassment in social VR effectively. Interviews lasted 102 minutes on average, and the longest one lasted almost 4 hours. Due to how in-depth and profound these interviews were, participants received a $50 Amazon digital gift card to compensate for their time after they completed the interviews.

Data Analysis. After the interviews were complete, recordings were first transcribed for further data analysis. We then utilized a thematic analysis approach [11, 12] to conduct an in-depth inductive qualitative analysis of the collected data. Following Braun and Clarke's [12] detailed guidelines and reproducible thematic analysis procedures, we analyzed all collected interview data in the following steps. (1) *Familiarizing ourselves with the data*: two authors closely read through the participants' transcribed narratives to identify information relevant to the research questions by highlighting them and taking notes to gain a full picture of participants' perceptions, expectations, and recommendations for leveraging AI companions to protect women from harassment in social VR. (2) *Generating initial codes*: The same authors began an iterative coding process. They independently assigned preliminary codes to identified information. Then, the two authors combined the identified codes, eliminated redundant codes, and ensured that highlighted content only aligned with a single code. For example, the quote *"I think an AI as a companion would help prevent harassment because most persons, when they notice the kind of AI you have, they become frightened by it"* was coded as "frightening presence," "effectively prevent harassment," which were then combined into "preventative interventions." (3) *Searching for themes*: These authors categorized codes into thematic topics related to our research questions and developed sub-themes from participants' descriptions. For example, codes about women's envisions of the AI companion's actions to discourage or deter their harassers in social VR were categorized as *Protective Companions*. (4) *Reviewing themes*: Two authors continued to discuss, integrate, and refine themes and sub-themes to streamline women's perceptions of AI companions in social VR to best capture and represent the data in relation to the research questions. (5) *Defining and naming themes*: All authors collaborated to refine these themes further and name the final set of themes. At this stage, all authors considered themes across the entire data set and identified the "essence" of what each theme is about [12]. (6) *Producing the report*: All authors selected the most compelling quotes as examples and logically drafted the structure of the findings. This phase aimed to create a narrative structure where all findings flowed naturally and coherently [12].

Positionality Statement. Due to the sensitive nature of our research focus, our identities and cultural backgrounds may influence our analysis and interpretation of the data [3, 43, 70]. Four out of five authors of this paper identify as straight and cisgender women, including three women of color. They also have extensive experiences in social VR both as women users and as researchers. Our own identities help us not only build in-depth understandings of the severe harassment risks women are facing in social VR but also better interpret women's own perceptions and envisions of leveraging AI companions to protect them from such harassment in future social VR spaces.

4 Findings

Overall, all 20 women considered the idea of designing AI companions as a new harassment mitigation approach to protect them in social VR novel and promising. They have collectively envisioned four nuanced ways in which AI companions can better help them navigate and manage harassment risks in social VR compared to existing safety mechanisms and strategies they have used (**RQ1**). Yet, they have also identified several potential challenges that need to be addressed when designing AI companions to protect women in future social VR spaces **(RQ2**). Table 2 summarizes our key findings.

4.1 Envisioning AI as Accessible Companions for Women Rather Than Depending on Any Human Companion's Availability or Power

Prior works have highlighted that while women social VR users' own harassment mitigation strategies, such as always hanging out with a group or relying on allies who are not considered marginalized in social VR (e.g., cis straight men), can be effective, these strategies place an excessive burden on women themselves and can further marginalize women in male-dominated social VR spaces [71]. In contrast, all our participants envision that AI has a significant comparative advantage over these existing strategies because AI can act as an **Accessible Companion** to women at any moment *rather than relying on any human companion's availability or power*, which can make women feel safer in any situation by being there when the user would otherwise be alone.

Indeed, the nature of AI technology itself as a computationally powerful system that can operate automatically and around the clock means that it is not subject to the same limitations as any human companions. In this sense, having an AI companion means that women social VR users could *access* the support and protection whenever and wherever they need in social VR spaces, rather than waiting for a human companion (e.g., a friend or an ally) to provide such support, not to mention that many women social VR users may not even have access to such human companions at all. P9 mentions, *"I get my AI to comfort me, I get my AI to be with me all the time."* Given that "not being alone" is widely considered an effective personal strategy to prevent harassment against women in social VR [71], AI as an Accessible Companion can be especially critical when women social VR users do not have access to other people to hang out with as a group, either because they have not made friends yet or because their social VR friends are unavailable at the time.

Additionally, having AI as an Accessible Companion could allow women users to feel safe when exploring new and unfamiliar situations and contexts in social VR. P15 explains this using the analogy of going to a party as a woman by herself where she does not know anyone,

"I don't know anyone there, why should I go? But sometimes your friend says, 'I'm going to go to that party. Would you come with me?' You say yes. Why? Because you know your friend is there, and even if you go to that party and you don't spend any time with your friend [...] going there with your friend makes you comfortable."

According to P15, an AI companion can act as a buffer to make women social VR users feel comfortable venturing into the unknown without feeling being alone, especially considering how pervasive harassment against women is in social VR spaces nowadays. For her, an AI companion is a "friend" who is always available and can always accompany her to a virtual party. P15's analogy also points out that while the AI companion's presence might be necessary for women's initial entrance into the unknown space, it is likely that women would not need the AI companion there with them once interactions have started to play out more, in the same way that you can go to a party with a friend and *"you don't spend any time with them."* In this sense, compared to women's existing safety strategy of "hanging out with a group," AI as an Accessible Companion provides similar benefits. Yet, it does not require women to actually find a human group or over-rely on AI all the time (e.g., once women feel comfortable, they will not need the AI companion with them).

Challenge to Maintain Social Connections: The Possibility of Diminishing Women's Connections with Other Humans in Social VR. However, some participants are deeply concerned that using AI as an Accessible Companion may make it difficult to form genuine human connections and friendships in social VR. This concern is driven by two factors related to an AI companion's automated and accessible nature.

First, participants like P11 believe that the automated nature of the protection provided by an AI companion might be *"too intrusive"* if it prevents women from interacting with others even in situations where women do not feel they are being harassed in social VR, such as: *"I fear it would cause way too many false positives and it could take away from the experience when exploring worlds"* (P11). In this sense, while our women participants acknowledge that mitigating harassment is an imperfect process regardless of who is doing the mitigating (e.g., *"I think an AI companion would work well, but don't forget, it won't halt it completely, neither would humans,"* P20), they are worried that AI companions may automatically block an otherwise friendly social VR user from interacting with them, limiting whom and how women could build connections within social VR. Moreover, some participants are worried that even the mere presence of an AI companion could dissuade other users from interacting with the user regardless of the AI companion's abilities. For example, P3 explains, *"Some who know the value [of the companion] or who fear harassment will treasure [it] [...] Others will feel like I'm a fraudster or I want to maybe cheat them or scam them or harass them."* Here, P3 points out that the purpose of the AI companion might not always be immediately distinguishable to other users, leading to misinterpretations of the user's intentions and motives. For P3, even by merely being present, an AI companion

Limitations of Women's Existing Safety Methods Identified in Prior Works	Women's Envisionings of AI Companions vs. Existing Safety Methods (RQ1)	Women's Perceived Challenges of AI Companions (RQ2)
Limitation 1: Reinforcing existing unequal power dynamics in social VR	As **Accessible Companions** to women rather than depending on any human companion's availability or power	Challenge to maintain women's social connections
Limitation 2: Offering limited informational, social, and emotional support to victims who have been harassed	1. As **Informational Companions** to effectively provide women with knowledge on how to deal with harassment	Challenge to build women's trust in AI-generated knowledge
	2. As **Emotional Support Companions** to help women recover from the internalized harm after being harassed in social VR	Challenge to provide women with a genuine sense of support
Limitation 3: Significantly reducing women's embodied experiences in social VR	As **Protective Companions** to proactively protect women from physicalized harassment while still ensuring their embodied experiences	Challenge to ensure appropriate usage of AI to protect, not harass, women

Table 2: Summary of Key Findings

could inadvertently hinder her ability to engage in friendly, desired interactions.

Second, some others are concerned that the AI companion would act as a convenient replacement for their human interaction (e.g., they will no longer need human companions to hang out with in social VR). For example, as a community leader in VRChat, P17 has extensive insight into how social VR can be a space where women who otherwise might not ever meet can find each other and form deeply personal connections through their embodied and immersive social interactions. Therefore, P17 is especially concerned that the potential use of AI companion that people can access anytime would lead to many social VR users, including women, forgoing human connection altogether because they may essentially rely too much on their AI companion for social connection and feel less motivated to seek out human connection in social VR: *"There's a lot of people that need other people and I think it's important for them to find them rather than having a fake person, a fake friend."*

In this sense, while the presence of an AI companion helps women not rely on any human companion's power to protect themselves, it may also make women social VR users feel like they do not need to seek out human social connections while in social VR, *"If you're looking for that social connection [from the AI companion], that's where it becomes problematic"* (P17). As such, designing AI as an Accessible Companion might both relieve women from unequal power dynamics when dealing with harassment against them and inadvertently reduce and even replace how they build connections with other human users in social VR spaces.

4.2 Envisioning AI as Informational Companions to Effectively Provide Women with Knowledge on How to Deal with Harassment

Although various platform-provided safety features (e.g., blocking, muting, reporting, and Horizon Worlds' space bubble feature [54]) have been introduced to help women social VR users combat harassment, all of our participants point out that no matter how many tools they have access to, it is difficult for them to manage harassment in social VR because no existing safety mechanism is sufficient to provide them with knowledge and/or knowledge recall on how to deal with harassment in-the-moment. As such, all participants welcomed the idea of AI as an **Informational Companion** by acting as a source of knowledge on how to deal with harassment when women need it.

For example, participants most often envision that AI as an Informational Companion would immediately tell or remind women users about the safety tools available to them after an incident of harassment takes place, *"And then they would probably also prompt you to like, 'Did you want to make a report about this?' And then it would lead you to where you would do that, so you wouldn't have to try to find it yourself"* (P8). Here, P8 believes that AI as informational companions to protect women should not only be reactive (e.g., reminding women of their options after an incident happens) but also proactive (e.g., actively guiding women to the option they choose). There is also a recognition amongst participants that such information should be accompanied by guidance on how to perform the suggested action. In this sense, compared to existing platform-provided safety features, AI as an Informational Companion both provides necessary in-the-moment knowledge and eases a woman user's mental load by eliminating the need to remember how to react to a harassment incident.

Additionally, as shown in prior works, anyone can be a harasser in social VR, including women themselves [27, 71]. Therefore, P4 sees AI as an Informational Companion as being especially well-positioned to push back against someone's own potentially harmful actions (including women themselves), which can help individuals be better community members in social VR: *"If you are trying to use harassing words, the AI could appear and try to caution [you] and also say the psychological effect and the negative aspect of what [you are] trying to do. Because at times, you may not really know."* Here, P4 re-envisions incorporating mental health information as informational support that helps everyone create a better environment for all,

recognizing that not every instance of harassment is necessarily intentional on the harasser's part.

Challenge to Build Human-AI Trust: Women's Uncertainties about the Credibility of AI-Provided Knowledge. However, our women participants also see a potential risk that should be addressed in designing and implementing AI as Informational Companions: how would women *trust* such knowledge provided by an AI companion, especially in a sensitive and highly personal context of harassment mitigation? P3 points out,*"my AI companion should have a good knowledge of a lot of issues happening in social VR because you can't lead me when you lack knowledge about something."* In this sense, women social VR users like P3 express concerns about the credibility and breadth of the AI's knowledge base regarding harassment in social VR, especially when women need to rely on such critical knowledge in a particularly vulnerable state (i.e., protecting themselves from harassment).

P17 reveals a similar sentiment, *"It's kind of hard because it's not necessarily a game where you can just put a guide on, it's not like WoW where you're like, "Oh, this is how you use your skills. This is how you open a menu," it would have to be like, 'How do I deal with people? How do I deal with the things that are going to come up in this game?'"* According to P17, while it might be acceptable for AI to provide women with instrumental knowledge about objective facts (e.g., *"how you open a menu"*), it is questionable to rely on AI-provided knowledge to deal with complex social issues in human interactions such as *"'How do I deal with people?"* For these women, due to the highly sensitive and complicated nature of harassment in social VR, asking women to depend on a technical entity (i.e., an AI) to deal with a social problem (i.e., harassment) may lead to several critical questions: should women trust such knowledge provide by an AI? To what degree should women follow AI-provided knowledge to deal with highly sensitive human interactions such as harassment in social VR? As such, how to foster women's trust in the credibility and quality of AI-provided knowledge regarding harassment mitigation is essential to designing future AI as informational companions whose intended benefits for protecting women can actually be accepted by women themselves.

4.3 Envisioning AI as Emotional Support Companions to Help Women Recover from the Internalized Harm after Being Harassed in Social VR

Harassment against women in social VR is well-recognized as particularly visceral, given the immersive and embodied nature of such attacks. Therefore, it is no surprise that women social VR users who must endure or witness this type of treatment are always searching for a source of comfort to help them feel better and recover from such traumatized experiences. Yet, women often lack effective strategy to overcome these internalized shame and fear stemming from being harassed in social VR, other than just building a "thick skin" [71]. Women social VR users also do not always have access in the moment to someone who can provide that comfort. Born from this need for a *"warm voice, a soft place"* (P1) after a harassing incident has occurred, participants envision AI as an **Emotional Support Companion** that they can immediately turn to for a sense of comfort and safety in cases of harassment in social VR. Many participants consider such support a much-needed yet critically lacking source for them to deal with harassment in social VR.

For example, several participants envision a system where a woman social VR user's AI companion detects harassment happening against them. It then automatically performs a combination of physical and verbal acts of comfort directed towards the user. In this sense, what makes an AI companion unique and most effective in helping women mitigate harassment in social VR is its more personal, intimate, and emotional nature. In doing so, the AI companion can support women social VR users in the moment *and* beyond the moment by providing words of comfort designed to invalidate whatever the harasser said about them and encouraging them to move on from the damaging harassment incident.

Additionally, participants envision emotional support coming from an AI companion for women to be focused more on uplifting women's self-esteem to counteract the negative impacts of harassment in social VR, as described by P12, *"I want it to look like me because most times whenever I see my avatar, it boosts my self-confidence. I see it and I'm like, 'Well, I'm actually not a bad human. I'm actually not disfigured. I'm actually beautiful.' So every time I get to see my AI companion [...] I feel like I do not need any other thing apart from who I am."* P12's statement paints a powerful picture of how emotional support aimed towards uplifting a woman's self-esteem can serve to counteract harmful thoughts that might become instilled in victims of harassment in social VR, e.g., being reminded that one is *"not a bad human"* or *"disfigured"* when being made fun of for their appearance.

Finally, some participants envision that the comforting power of an AI companion could be especially enhanced through avatar customization that emphasizes friendliness and companionship, such as through the appearance and behavior of a pet (e.g., *"a fox, or a dog, or a cat to be your little animal AI presenting thing,"* P8) or anthropomorphized inanimate object (e.g., *"something like Clippy,"* [3] P8). This preference makes sense in light of the uncomfortable "uncanny valley" [55] effect that humanoid avatars can have on users. P8 further explains, *"Something nice and friendly-looking [...] because of it's too humanoid, it can turn into the uncanny valley type of thing and people would just be kind of weary of it."* Therefore, while some participants like P12 might find more comfort in a humanoid avatar designed to uplift their self-esteem, others like P8 might find a humanoid avatar too off-putting to seek comfort from, further emphasizing the importance of AI companion avatar customization to fit individual comfort needs.

Challenge to Create a Genuine Feeling of Support: The Inherent Conflict between the Social Nature of Care and Comfort and AI as an Artificial Computer Program. Our women participants especially appreciate the more personal and intimate nature of AI as an Emotional Support Companion to help them recover from the severe damage of being harassed in social VR. However, they also express concerns about how existing AI technologies lack the necessary components to create a genuine feeling of supportive friendship or emotional bond.

[3]Here P8 refers to Microsoft's Clippit, a cartoon paperclip with human-like features used to assist pre-2007 Microsoft Office users.

For example, some women doubt the idea of forming a friend-like bond with an AI entity for emotional support, care, and comfort especially in a highly sensitive and personal context like harassment mitigation in social VR, regardless of how friend-like the design of an AI companion is. For these participants, the social nature of the sense of care and comfort fundamentally depends on their relationship partner being a human being, which is unachievable by the very nature of AI as an artificial computer program. This is likely the case for many participants because their prior experiences with technologies equivalent to AI companions (e.g., Siri and Alexa) lead them to view AI as lacking the ability to emulate the human-like qualities needed for actually comforting harassment victims after such incidents (e.g., *"I'm in the camp that AI isn't sentient,"* P19).

More importantly, many of our participants view AI not as a social and emotional entity at all but as a tool to be used, such as:

"I would view the AI as a tool to help me mitigate harassment since it's not a real person, I can't really connect to an AI that well [...] It would also feel more natural talking/getting protected to a human instead of an AI." (P7)

"I just don't really put a face to many AIs, I guess [...] I know with a lot of AI programs, like Alexa or Siri, I know a lot of them tend to use female voices because people tend to find female voices more soothing or comforting, and that sells more to them. I do know something about the whole identification with AI thing, but for me personally, I don't really care if the AI is what it is I guess. So, it's just a tool to me." (P18)

These participants underscore how women's view of AI's nature undermines an AI companion's ability to effectively help them recover from the internalized harm after being harassed. For them, creating a genuine feeling of such highly personalized support in a sensitive context would involve more emotional bonding than what an AI companion's essential role as an instrumental tool could currently provide.

4.4 Envisioning AI as Protective Companions to Proactively Protect Women from Physicalized Harassment While Still Ensuring Their Embodied Experiences in Social VR

Lastly, almost all participants envision AI as **protective companions** to protect them from physicalized harassment while still allowing them to fully engage in embodied interactions in social VR – e.g., to let them safely engage in social VR without putting on personal space bubbles or self-disguise. They especially highlight two ways in which such AI companions are considered "proactive": (1) providing preventative interventions; and (2) providing active interventions.

First, *preventative interventions* can best be thought of as (in)actions by the AI companion that could deter harassers or prevent harassment from reaching women social VR users in the first place. Several participants like P5 explain that the mere presence of AI as proactive companions can serve as a signal to others not to mess with the woman user,*"I think an AI as a companion would help prevent harassment because most persons, when they notice the kind of AI you have, they become frightened by it"* (P5).

Interestingly, our women participants who felt that the mere presence of AI as protective companions could serve as a preventative measure expressed preferences for masculine or aggressive-looking avatar customization options, as they associate these types of avatars with protection and intimidation. Participants like P14 and P5 envision male-presenting AI companion avatars as being the most likely to deter potential harassers because, as P5 explains, *"People will respect the fact that it's a male AI. They might be frightened by it"* (P5). This sentiment echoes prior works' findings on how women social VR users will sometimes use proximity to a male-presenting friend to feel safe from harassment [71]. However, the intimidating appearance must still be balanced in a way that makes it pleasant for the individual user to be around while still scaring away potential harassers targeting women. P6 summarizes, *"Having something that can be comforting and small like a house cat is ideal for most situations to me, but in a situation where I don't feel safe I would prefer something bigger and more protective."* P6's statement clearly underscores our participants' beliefs that AI companions to protect women social VR users should respond to the woman's dynamic needs for protection at the moment, e.g., appearing intimidating only *"in a situation where I don't feel safe."* In this way, an intimidating appearance can act as a preventative and a reactive action depending on the woman user's preferences without limiting or interrupting how she would interact with others.

Second, *active interventions* can best be described as actions that the AI companion performs that are directed against women's harasser, such as: *"maybe it [the AI companion] could detect abusive words and have the potential to inflict pain... or sanctions for the harasser"* (P2). For instance, P13 envisions AI as a Protective Companion putting their in-world body in between her harasser and herself to prevent the harasser from escalating their behavior through a physical attack, just like a bodyguard would. P10 further envisions the AI companion speaking up to defend a woman social VR user by verbally warning the harasser to watch their behavior (e.g., *"Harassment detected. Choose your words carefully. Or stop your behavior."*), which can be especially helpful for women social VR users who do not feel confident confronting others.

Some participants even advocate for the AI companion to have the ability to remove the harasser from the presence of the woman victim physically, especially when sexual harassment is involved. P13 envisions that the AI companion would physically remove the harasser from the in-world location altogether, e.g., by walking them out of the specific social VR room. Further, P9 reveals that the AI companion would essentially scare away harassers by chasing them until they are out of the purview of the woman user (*"chase away users that are intending to harass me or humiliate me"*). In a final escalation, P14 even shares that an AI companion would use physical force to remove the harasser rather than using calmer (e.g., walking them away) or no-contact methods (e.g., chasing away), *"Defend me when I have attackers. To drive away harassers from my space. Grab my harasser's shirt and pull him off."* While this envisioning of AI as protective companions for women might appear to be an escalation of aggression, P14 seems to advocate this method when women encounter ongoing, extremely severe physical harassment in social VR (e.g., sexual assaults).

Challenge to Ensure Appropriate Usage: The Potential of AI Becoming a New Tool for Harassing Women. Yet, our

women participants also express several concerns about how AI as protective companions may become a new tool to bully and harass them if not being used appropriately with restrictions. For example, many of our participants feel that the use of such AI companions might alienate them from the larger social VR community and even make them targets for harassment. P18 reveals, *"I feel if somebody saw that you had an AI companion with you to prevent harassment, people are just going to tell you that you're a pussy."* As such, participants like P16 believe that depending on AI as protective companions for women may *"kind of undermine your own authority to deal with the problem"* by intervening in various ways on women's behalf, which actually makes women less able to learn to deal with harassment on their own in the long run.

Others also point out that due to the physicalized nature of such protective AI companions' potential actions, they can be misused to bully and harass women more, rather than protecting them. P16 explains, *"It wouldn't have much authority of its own to do anything. If it did, that would be a problem in and of itself. So let's say my AI companion can kick someone out of the app, well, that's not fair. What if someone just decides to be a dick and start kicking people out of the app?"* For P16, any intervention action that directly affects another user's ability to engage in social VR generally, rather than just affecting their ability to interact with the user requesting protection (e.g., blocking the harasser from the victim's view), would be considered inequitable.

This perception of inequity is likely rooted in the fact that such an action crosses a line between protection to keep a woman victim safe and unilateral punitive punishment of a harasser's actions by stripping bodily and experiential autonomy from said user. Indeed, while some social VR platforms have integrated tools that allow users to remove other users from social VR environments, these come with specific restrictions. For example, Horizon Worlds allows communities to expel users from a space, but only if most users within that space vote affirmatively to do so [54]. These restrictions likely result from the recognition that giving this kind of power to individual users, such as via an AI companion, can make the AI companion a *new tool for harassing*. As such, although participants like P16 believe that AI as protective companions could effectively shield women from physicalized sexual harassment without limiting their own embodied interactions, they warn that designing such AI companions must balance everyone's equal access to bodily autonomy to mitigate the potential harm of misuse.

5 Discussion

To further fulfill people's needs for companionship in the modern digital society, AI companions exist in many personalized forms to help foster a sense of safety, support, and comfort across various digital landscapes [2, 16, 38, 61, 62, 83, 92]. Social VR platforms, however, do not at the moment provide users with any tools to create or interact with an AI companion of any kind. This in turn allows our women participants to engage in their own ways with the idea of how an AI companion can be leveraged to help women better mitigate the severe harassment risks in social VR they have been dealing with. Grounded in our key findings shown in Table 2, in this section, we first discuss how our findings offer new insights to innovate existing AI-based harassment mitigation approaches to better protect marginalized users such as women in social VR. Built upon women's own understandings of inherent challenges in leveraging such AI companions to protect them, we then reflect on sociotechnical complexities regarding creating effective and sensitive AI solutions for social VR environments, which must be carefully considered when designing and implementing AI companions in future social VR spaces. Informed by these insights, we also propose three high-level principles aimed at designing future AI companions as a new harassment mitigation approach to protect women and other marginalized social VR users.

5.1 Innovating How We Can Better Protect Women in Social VR Through AI

One of the most important highlights from our findings is that all of our women participants enthusiastically envision future AI companions as a new approach to innovate how women can be better protected, supported, and cared for in social VR, especially when they have to deal with severe harassment risks against them in these spaces. Our work reveals several new ways in which advanced AI technologies are envisioned by women themselves to help mitigate social VR harassment targeting women in more nuanced ways in comparison to both (1) *existing AI-based harassment mitigation methods* [22–24, 72] and (2) *safety mechanisms and strategies women have used in social VR* [8, 9, 27, 57, 58, 71, 76].

First, compared to existing AI-based harassment mitigation methods in social VR, AI companions innovate how women can be better protected from harassment by acting beyond a *moderator* or a *punisher*. Indeed, prior literature has explored the prospect of leveraging AI in social VR for harassment mitigation, especially in the forms of AI-based moderation [72], AI-powered moderator avatars to protect children [23, 24], and NPCs as mediators to intervene in emerging safety risks [93]. These works seem to highlight AI's main roles as either a *moderator*, who can respond dynamically and specifically based on individual users' and communities' conceptions of harassment (e.g., [72]), or a *punisher* who can directly intervene harassment incidents and stop harassers [22–24, 93]. In contrast, our women participants clearly envision that using AI to protect women from harassment in social VR should go beyond just a figure for punishment or a system to merely monitor, detect, and correct people's behaviors. Rather, they highlight the urgently needed efforts to design AI companions based on many of women's under-addressed harassment mitigation needs *before*, *during*, and *after* a harassment incident. Therefore, they have envisioned AI companions to make them appear with others without any actual human companions to prevent harassment targeting alone women (*Accessible Companions*), to help them learn how to deal with harassment both before any incident or during an ongoing incident (*Informational Companions*), to heal their internalized damage after being harassed in social VR (*Emotional Support Companions*), and to take proactive actions to protect them from ongoing physicalized harassment (*Protective Companions*). These women's collective voice is clear: only using AI to moderate and intervene is insufficient to foster safe social VR environments for marginalized users who are vulnerable and traumatized by harassment in social VR. As these women have highlighted, while it is important to design AI companions that look big and intimidating

like a bodyguard to provide them with preventative and active protections, it is equally crucial to design AI companions that appear to be comforting and small like a house cat so that they could feel safe, supported, and not alone in social VR spaces.

Second, compared to safety mechanisms and strategies women have used in social VR, AI companions innovate how women can be better protected from harassment by providing valuable alternatives to enhance their benefits while reducing limitations. Existing literature has clearly shown that while acknowledging the effectiveness of existing safety mechanisms and strategies they have used in social VR, women have also shared deep concerns about various limitations of these tactics and expressed their strong desires to seek potential alternatives [8, 9, 27, 57, 58, 71, 76]. In this sense, our findings indeed provide empirical evidence of how women envision AI companions to help both enhance the benefits of their own harassment mitigation tactics and reduce the identified limitations in three ways, as shown in Table 2. Above all, by highlighting the importance of designing AI as *Accessible Companions*, women believe that AI companions could help them continue to leverage their successful strategies of "hanging out with a group" and "having connections with other allies" [27, 71] but without further subjecting women to existing unequal power dynamics in social VR (e.g., women must depend on men's availability and power to make themselves safe), which addresses *Limitation 1*. By highlighting the importance of designing AI as *Informational Companions* and *Emotional Support Companions*, women envision that AI companions could directly fill a gap by providing them with the much-needed informational, social, and emotional support that their existing harassment mitigation methods largely fail to offer [8, 9, 27, 57, 58, 71, 76], which addresses *Limitation 2*. Further, by highlighting the importance of designing AI as *Protective Companions*, women expect that AI companions could help them fully engage in embodied interactions and safely present their identities (i.e., as women) in social VR by shielding them from physicalized harassment without the need for using personal space bubbles to distance themselves from everyone's touch or self-disguise to hide their own identity [8, 9, 27, 57, 58, 71, 76], which addresses *Limitation 3*.

5.2 Sociotechnical Complexities for Leveraging AI Companions to Protect Women in Social VR

While women social VR users' own voices and perceptions clearly paint an overly positive image of AI companions as a new and promising harassment mitigation approach in social VR, they still express their profound concerns about several challenges in this new approach (Table 2). In this section, we further reflect upon various sociotechnical complexities for designing and implementing AI companions to protect women in future social VR spaces.

Above all, creating an AI companion that is capable of providing women with context-sensitive emotional support while also understanding the nuanced cultural and gender dynamics involved in mitigating online harassment against women presents significant technical and design complexities. Although previous works have proposed several important principles to design human-AI companionship in general, such as Strohmann et al.'s design theory for virtual companionship [82], our participants envision and emphasize the potential of using AI companion in a sociocultural sensitive context to help women navigate and manage harassment risks in social VR. Therefore, designing such AI companions should go beyond general principles such as human-like design, adaptivity and adaptability, and proactivity and reciprocity [65, 82] but focus more on women's nuanced social and emotional needs when dealing with social VR harassment as a new context for human-AI companionship.

For example, our participants envision that such AI companions should be accessible to women in any occasions where they need it; yet it can also be disabled to prevent overshadowing their connections with other social VR users. Such AI companions should provide women with the necessary knowledge about how to deal with harassment whenever women encounter it; yet it should also let women verify and assess the credibility of this AI-generated information. In this sense, how can AI companions be programmed to understand and immediately respond to these varied social interactions women may encounter within social VR platforms? And what specific protocols or guidelines need to be established to ensure that AI companions work in conjunction with human moderators and existing content moderation tools (e.g., verifying AI-generated information and recommendations)? Addressing these questions would be important for social VR researchers and developers to practically integrating potential AI companion systems within current social VR frameworks, which we also further discuss in section 5.3.

In addition to these technical and design complexities, a more fundamental and ethical question is: *should an AI companion be designed as a technical solution or a social entity when it comes to protecting women from harassment in social VR?* Existing definitions of AI companions often highlight that such companions should establish and maintain long-term social bonds with their users, which makes them "companions" rather than "assistants" [41, 65, 82]. As such, this focus comes with various ethical risks that must be considered when establishing human-AI companionship. These include privacy and security issues, surveillance and censorship (e.g., inappropriately collecting behavioral and interaction data to learn the user's behavior [4, 35, 39]), and loss of user autonomy [75]. Further, robot and AI companions may risk dehumanizing and isolating users, especially those belonging to marginalized and vulnerable groups, from critically needed human companions [6]. Anthropomorphized AI companions may even lead to the ethical risk of human users inappropriately forming emotional attachments with AI, which may replace their friendships with humans [13] and lead to misplaced feelings of trust [64]. As a result, AI companions may make these individuals become more isolated [4, 69].

In our work, women social VR users further emphasize these ethical complexities when envisioning using AI companions to protect them in social VR. On the one hand, they clearly envision that their AI companion should go far beyond a technical solution like an advanced computer program that mainly focuses on automatically modulating and detecting. For them, this is insufficient to protect them and make them feel safe. In contrast, they show strong desires for an AI companion as a social entity to meet their various social and emotional needs due to the highly sensitive, subjective, and

personalized nature of harassment in social VR. On the other hand, designing AI companions as social entities also creates nuanced ethical issues that might actually harm women: Will AI companions replace women's connections with their actual human friends in and out of social VR who can help them move on from the harmful experience and to counteract any damage such harassment might do to a woman's self-esteem? Even if an AI companion can provide critically needed emotional support after a woman is harassed in social VR, is it actually genuine or meaningful to the victim because it is not even from a human?

Taken together, our study also calls for future research to reflect upon how to strategically navigate various technical, design, and ethical complexities when designing AI companions to fulfill women social VR users' multidimensional harassment mitigation needs without harming their self-agency, human relationships, and supportive networks. We discuss this in the next section.

5.3 Designing Future AI Companions to Protect Women and Other Marginalized Users in Social VR

Built upon our reflections detailed above, we propose three high-level principles for approaching future AI companion design to protect women and other marginalized users in social VR spaces. Designing such AI companions does not only involve navigating complex technical design but also requires careful ethical considerations. Therefore, rather than focusing on proposing generic new design features, we view these principles as a foundation to open up conversations, and help us rethink more nuanced harassment mitigation strategies in social VR spaces. Additionally, these principles do not aim to replace or contradict HCI and social VR researchers' current focus on designing new safety features for harassment mitigation in social VR (e.g., AI-based moderation). Instead, we hope these principles may complement these efforts by specifically focusing on marginalized social VR users, such as women's, under-addressed harassment mitigation needs.

Principle 1: Designing AI companions based on women's self-agency to foster mutual support, emotional bonding, and trust among marginalized users rather than replacing their human connections. Indeed, stemming from our women social VR users' struggles with treating AI companions as computer programs vs. social entities, they warn of the danger of reducing women's connections with other human friends and allies in social VR and subjecting women to a "delusion" of being supported and cared with little genuine social value (e.g., being supported by an artificial computer program rather than a real human). In this sense, women do hope to leverage advanced AI technologies such as AI companions to better help them mitigate damages of social VR harassment socially and emotionally rather than just instrumentally (e.g., moderating and punishing). To achieve this goal, first, AI companion should be designed in a way to *support*, rather than *surpass*, women social VR users' self-agency [28, 52]. This means that women social VR users would have complete agency to decide *when and how they wish to use AI companions for harassment mitigation.* For example, the AI companion would work to learn how the woman user defines harassment and what type of support they would prefer to receive (e.g., asking a nearby woman friend for help) by observing when the user calls upon them to help and/or when the user executes the advised harassment mitigation strategy. As such, the AI companion does not need to be programmed to have a one-size-fits-all framework to respond to all the different social interactions women may encounter in social VR. Rather, women would craft their relationship with the AI companion and define what they would need the AI companion for and under which situations.

Built upon this self-agency, second, such AI companions can be designed to help women seek mutual support and emotional bonding with other marginalized users (e.g., other women) rather than completely replacing their human connections. For example, similar to Lyft's Women+ Connect feature [47], an AI companion can be configured to help a woman connect with other women users and build a sense of community if they do not want to hang out alone in social VR, which also relieves women from the excessive burden of seeking and maintaining such allies and the need to depend on men. This approach may simultaneously help women build trust on knowledge and information regarding harassment mitigation provided by the AI companion, as they would be able to confirm such information with other women who have similar experiences. Certainly, when other women are not available or around, women social VR users would still be able to use their AI companions for protection and support. However, the main goal of such companions should be enhancing and facilitating how women can be connected with and support each other, not replacing such peer support.

Principle 2: Designing AI companions to enhance a sense of comfort rather than alienation. Many of our women participants also feel that using AI companions might alienate them from the larger social VR community because they may make other social VR users around them feel uncomfortable and anxious. Worse still, rather than serving as a mechanism to intimidate and deter harassers, our participants further point out that the mere presence of a physicalized AI companion can make them a *more visible target for harassment* in ways reminiscent of how embodied avatars can make marginalized users such as women more visible targets for harassment [27, 72]. Therefore, there seems to be a strong belief that AI companions should be designed to enhance women's sense of comfort rather than alienation, mainly through how often they wish the AI companion to appear and more nuanced customization of such companions' appearances.

Given our findings, we first suggest that women should be able to choose between different options of how often they wish the AI companion to appear to them (e.g., "Always be with me," "Only appears when I call it," "Only appears when it detects harassment," and/or "Never appears to me."). Then, we suggest designing for AI companion appearance customization should go beyond typical avatar customization (e.g., different genders, ethnicities, and body types for humanoid avatars). Instead, such design would necessarily need to include (1) options for pet or anthropomorphic object-like avatars to fulfill emotional support-type comfort needs addressed in 4.3; and (2) options for physically imposing humanoid avatars to fulfill security-type comfort needs addressed in 4.4. And more critically, we also suggest that customization of the AI companion's appearance must necessarily include certain feature(s) that give users the ability to gain access to the benefits of the AI companion

(e.g., its reassuring presence as a protector like "Big Buddy" [23, 24]) without also making that user a more visible target for harassment or alienation. We envision several potential avenues for achieving this, including but not limited to an option to "hide" the companion's avatar from everyone's view (to avoid targeting) except the user's own (to give visual reassurance to the user); or an evolving avatar that has an innocuous or harmless appearance (e.g., a house cat) most times, which will not intimidate potential friends but can transform into a more powerful form when it detects harassment (as explained by P6 in 4.4) to act as a **preventative** mechanism.

Principle 3: Designing AI companions with careful consideration of unintended use to prevent personal abuses. Finally, our women participants share alarming concerns about the prospect of abusing AI companions as a new tool for harassing women and other marginalized communities (e.g., through the AI companion's active interventions). Indeed, our participants envision the AI to serve the sole interest of one and only one individual, even in potentially violent ways (e.g., physically pushing away other social VR users to protect a woman user). Given this, we suggest that the design of future AI companions must necessarily involve the careful consideration of all the ways in which an individual might abuse an AI companion if given the opportunity.

For example, one suggestion may be that future design of AI companions for harassment mitigation in social VR should not incorporate the use of most, if not all, physical violence (e.g., physically attack, touch, grab, or otherwise interfere with another user), as their use is likely to cause a high level of distrust amongst all social VR users in the AI companion system as a whole and/or could be used as a tool for harassing rather than protecting marginalized social VR users (e.g., commanding one's AI companion to sexually harass a woman). That is not to say that women's needs for AI as protective companions are invalid or unimportant to address. Rather, we suggest that future designers of AI companions should take into consideration if the need for using AI companions as personalized protectors or bodyguards might be better achieved together with other existing harassment mitigation strategies (e.g., human moderators [9, 67], AI-based moderation systems [22–24, 72], or social VR consent mechanics [73, 95]). For instance, it is essential to provide well-defined guidelines and community training on appropriate ways to use AI companions to defend oneself (e.g., do not physically attack a harasser to stop their ongoing harassing behavior) and how to leverage existing content moderation infrastructure in social VR to manage inappropriate uses of AI companions (e.g., report the abuse to AI moderator or a nearby human moderator). These efforts would help maximize the benefits of AI companions for protecting marginalized users like women while minimizing the severe and potentially damaging risks of abusing such technologies.

5.4 Limitations

We acknowledge that our recruitment methods may have led to potential self-selection bias, e.g., only social VR users who are also active social media users may have responded. However, the individuals recruited through these methods provide unique insights into social VR users' needs and the envisions of AI companions for harassment mitigation that are much needed for HCI and social VR research. Additionally, despite our efforts to recruit women social VR users across various countries and cultures and having a highly diverse sample regarding ethnicity, most participants are located in the USA. As harassment is a culturally contextualized construct, it is important to recruit participants who speak other languages or are from non-Western cultures in future studies to better understand women social VR users' perceptions of both opportunities and challenges of leveraging AI companions to protect them in social VR spaces. It is also important to note that although some of our participants have occupations or are in schoolwork related to a technology sector, most participants do not have any specific experience building or developing AI technology. Given this, our participants' envisions do not particularly focus on technical elements of how to build AI companions in social VR to protect them but rather reflect how their personal perceptions of and experiences in social VR combine with prior experiences with digital companions or AI companions in other online contexts (e.g., chatbots or NPCs). As discussed in this paper, creating such effective and sensitive AI companions and practically integrating them within current social VR environments would need to navigate various technical and design challenges, which would require more future work.

6 Conclusion

With growing concerns over harassment against women in embodied social VR spaces [7–9, 27, 57, 58, 71, 80, 85], in this work, we turn an eye towards an understudied lens (i.e., **AI companions**) and have revealed women social VR users' own envisionings for leveraging AI as *Accessible Companions*, *Informational Companions*, *Emotional Support Companions*, and *Protective Companions* to better protect them. Based on women's own concerns about several challenges in this new approach, we have also reflected upon various sociotechnical complexities for designing and implementing such AI companions and propose three high-level design principles to inform future efforts to create AI companions as a new approach for harassment mitigation in social VR. In doing so, we hope to contribute to ongoing discussions on better fulfilling marginalized social VR users', such as women's, under-addressed harassment mitigation needs without harming their self-agency, human relationships, and supportive networks.

Acknowledgments

We thank our participants and the anonymous reviewers. This work was supported by the National Science Foundation awards # 2112878 and #2342393.

References

[1] Carolina Are. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies* 20, 5 (2020), 741–744.
[2] Margaret Arnd-Caddigan. 2015. Sherry Turkle: Alone Together: Why We Expect More from Technology and Less from Each Other. *Clinical Social Work Journal* 43, 2 (June 2015), 247–248. doi:10.1007/s10615-014-0511-4
[3] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 675–684.
[4] Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. *An introduction to ethics in robotics and AI.* Springer Nature.
[5] Alan M Beck and Aaron Honori Katcher. 1996. *Between pets and people: The importance of animal companionship.* Purdue University Press.
[6] Andrea Bertolini and Giuseppe Aiello. 2018. Robot companions: A legal and ethical analysis. *The Information Society* 34, 3 (2018), 130–140.

[7] 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.

[8] 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.

[9] 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.

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

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[13] Elizabeth Broadbent. 2017. Interactions with robots: The truths we reveal about ourselves. *Annual review of psychology* 68 (2017), 627–652.

[14] William M Bukowski, Betsy Hoza, and Michel Boivin. 1993. Popularity, friendship, and emotional adjustment during early adolescence. *New directions for child and adolescent development* 1993, 60 (1993), 23–37.

[15] Tara Capel, Bernd Ploderer, Filip Bircanin, Simon Hanmer, Jamie Paige Yates, Jiaxuan Wang, Kai Ling Khor, Tuck Wah Leong, Greg Wadley, and Michelle Newcomb. 2024. Studying Self-Care with Generative AI Tools: Lessons for Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1620–1637.

[16] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Designing Interactive Systems Conference 2021*. 1817–1833.

[17] 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.

[18] 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

[19] 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.

[20] Katharina Emmerich, Patrizia Ring, and Maic Masuch. 2018. I'm glad you are on my side: How to design compelling game companions. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 141–152.

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

[22] Cristina Fiani, Robin Bretin, Shaun Alexander Macdonald, Mohamed Khamis, and Mark Mcgill. 2024. " Pikachu would electrocute people who are misbehaving": Expert, Guardian and Child Perspectives on Automated Embodied Moderators for Safeguarding Children in Social Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.

[23] Cristina Fiani, Robin Bretin, Mark McGill, and Mohamed Khamis. 2023. Big Buddy: A Simulated Embodied Moderating System to Mitigate Children's Reaction to Provocative Situations within Social Virtual Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[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] 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.

[26] Guo Freeman and Donghee Yvette Wohn. 2020. Streaming your identity: Navigating the presentation of gender and sexuality through live streaming. *Computer Supported Cooperative Work (CSCW)* 29, 6 (2020), 795–825.

[27] 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.

[28] Adina Friedman and Jacob Schrum. 2019. Desirable behaviors for companion bots in first-person shooters. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[29] 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.

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

[31] HBO. 2022. We met in virtual reality. https://www.hbo.com/movies/we-met-in-virtual-reality.

[32] 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)*.

[33] Christopher J. Headleand, James Jackson, Ben Williams, Lee Priday, William J. Teahan, and LLyr Ap Cenydd. 2016. How the Perceived Identity of a NPC Companion Influences Player Behavior. In *Transactions on Computational Science XXVIII*, Marina L. Gavrilova, C.J. Kenneth Tan, and Alexei Sourin (Eds.). Vol. 9590. Springer Berlin Heidelberg, Berlin, Heidelberg, 88–107. doi:10.1007/978-3-662-53090-0_5 Series Title: Lecture Notes in Computer Science.

[34] Eva V Hoff. 2004. A friend living inside me—The forms and functions of imaginary companions. *Imagination, Cognition and Personality* 24, 2 (2004), 151–189.

[35] Andreas Huber, Astrid Weiss, and Marjo Rauhala. 2016. The ethical risk of attachment how to identify, investigate and predict potential ethical risks in the development of social companion robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 367–374. doi:10.1109/HRI.2016.7451774

[36] Lynn Jamieson. 2013. Personal relationships, intimacy and the self in a mediated and global digital age. In *Digital sociology: Critical perspectives*. Springer, 13–33.

[37] 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.

[38] Bijan Khosrawi-Rad, Ricarda Schlimbach, Timo Strohmann, and Susanne Robra-Bissantz. 2022. DESIGN KNOWLEDGE FOR VIRTUAL LEARNING COMPANIONS. *Proceedings of the 2022 AIS SIGED International Conference on Information Systems Education and Research* (Jan. 2022). https://aisel.aisnet.org/siged2022/6

[39] Jacqueline M. Kory-Westlund and Cynthia Breazeal. 2019. A Long-Term Study of Young Children's Rapport, Social Emulation, and Language Learning With a Peer-Like Robot Playmate in Preschool. *Frontiers in Robotics and AI* 6 (2019). doi:10.3389/frobt.2019.00081

[40] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[42] Michael Sangyeob Lee and Carrie Heeter. 2012. What do you mean by believable characters?: The effect of character rating and hostility on the perception of character believability. *Journal of Gaming & Virtual Worlds* 4, 1 (2012), 81–97.

[43] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.

[44] Mei Yii Lim. 2012. Memory models for intelligent social companions. In *Human-computer interaction: The agency perspective*. Springer, 241–262.

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

[46] Emma J Llansó. 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society* 7, 1 (2020), 2053951720920686.

[47] Lyft. 2024. Introducing Women+ Connect. https://www.lyft.com/women+.

[48] Daniel Madden and Casper Harteveld. 2021. "Constant Pressure of Having to Perform": Exploring Player Health Concerns in Esports. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[49] Daniel Madden, Yuxuan Liu, Haowei Yu, Mustafa Feyyaz Sonbudak, Giovanni M Troiano, and Casper Harteveld. 2021. "Why Are You Playing Games? You Are a Girl!": Exploring Gender Biases in Esports. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[50] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. It is complicated: Interacting with children in social virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 343–347.

[51] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 472–483.

[52] Kevin McGee and Aswin Thomas Abraham. 2010. Real-time team-mate AI in games: a definition, survey, & critique. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*. ACM, Monterey California, 124–131. doi:10.1145/1822348.1822365

[53] Joshua McVeigh-Schultz, Elena Márquez Segura, Nick Merrill, and Katherine Isbister. 2018. What's It Mean to" Be Social" in VR? Mapping the Social VR Design Ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 289–294.

[54] Meta. 2022. Safety and privacy in Meta Horizon Worlds. https://www.meta.com/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/index-safety-and-privacy/.

[55] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[56] Yifat Nahmias and Maayan Perel. 2021. The oversight of content moderation by AI: impact assessments and their limitations. *Harv. J. on Legis.* 58 (2021), 145.

[57] 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.

[58] 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.

[59] Benjamin Paaßen, Thekla Morgenroth, and Michelle Stratemeyer. 2017. What is a true gamer? The male gamer stereotype and the marginalization of women in video game culture. *Sex Roles* 76, 7 (2017), 421–435.

[60] Bianca Pani, Joseph Crawford, and Kelly-Ann Allen. 2024. Can generative artificial intelligence foster belongingness, social support, and reduce loneliness? A conceptual analysis. *Applications of Generative AI* (2024), 261–276.

[61] Maria Perez-Ortiz, Claire Dormann, Yvonne Rogers, Sahan Bulathwela, Stefan Kreitmayer, Emine Yilmaz, Richard Noss, and John Shawe-Taylor. 2021. X5Learn: A Personalised Learning Companion at the Intersection of AI and HCI. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 70–74. doi:10.1145/3397482.3450721

[62] Zahy Ramadan, Maya F. Farah, and Lea El Essrawi. 2021. From Amazon.com to Amazon.love: How Alexa is redefining companionship and interdependence for people with special needs. *Psychology & Marketing* 38, 4 (2021), 596–609. doi:10.1002/mar.21441 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21441.

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

[64] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

[65] Ayanda Rogge. 2023. Defining, designing and distinguishing artificial companions: A systematic literature review. *International Journal of Social Robotics* 15, 9 (2023), 1557–1579.

[66] Karen S Rook. 1987. Social support versus companionship: effects on life stress, loneliness, and evaluations by others. *Journal of personality and social psychology* 52, 6 (1987), 1132.

[67] 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.

[68] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. " They Don't Leave Us Alone Anywhere We Go" Gender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[69] Matthias Scheutz. 2011. 13 The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. *Robot ethics: The ethical and social implications of robotics* (2011), 205.

[70] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 5412–5427.

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

[72] 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.

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

[74] Gavin Scott and Foaad Khosmood. 2018. A Framework for Complementary Companion Character Behavior in Video Games. http://arxiv.org/abs/1808.09079 arXiv:1808.09079 [cs].

[75] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology* 14 (2012), 27–40.

[76] 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.

[77] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601.

[78] 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.

[79] 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.

[80] 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/.

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

[82] Timo Strohmann, Dominik Siemon, Bijan Khosrawi-Rad, and Susanne Robra-Bissantz. 2023. Toward a design theory for virtual companionship. *Human–Computer Interaction* 38, 3-4 (2023), 194–234.

[83] Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of Medical Internet Research* 22, 3 (March 2020), e16235. doi:10.2196/16235

[84] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.

[85] Chris Vallance. 2024. Police investigate virtual sex assault on girl's avatar. hthttps://www.bbc.com/news/technology-67865327.

[86] Aditya Vashistha, Abhinav Garg, Richard Anderson, and Agha Ali Raza. 2019. Threats, abuses, flirting, and blackmail: Gender inequity in social media voice forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[87] 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.

[88] Vogels, Emily A. 2021. *The State of Online Harassment*. Technical Report. Pew Research Center, Washington, D.C. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[89] Amy K Way, Robin Kanak Zwier, and Sarah J Tracy. 2015. Dialogic interviewing and flickers of transformation: An examination and delineation of interactional strategies that promote participant self-reflexivity. *Qualitative Inquiry* 21, 8 (2015), 720–731.

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

[91] Georgios N. Yannakakis and Julian Togelius. 2018. Playing Games. In *Artificial Intelligence and Games*, Georgios N. Yannakakis and Julian Togelius (Eds.). Springer International Publishing, Cham, 91–150. doi:10.1007/978-3-319-63519-4_3

[92] Han Yu, Chunyan Miao, Cyril Leung, and Timothy John White. 2017. Towards AI-powered personalization in MOOC learning. *npj Science of Learning* 2, 1 (Dec. 2017), 1–5. doi:10.1038/s41539-017-0016-3 Number: 1 Publisher: Nature Publishing Group.

[93] 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.

[94] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

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