Harassment in Social Virtual Reality: Challenges for Platform Governance

LINDSAY BLACKWELL, Oculus VR, USA
NICOLE ELLISON, Oculus VR, USA
NATASHA ELLIOTT-DEFLO, Oculus VR, USA
RAZ SCHWARTZ, Oculus VR, USA

In immersive virtual reality (VR) environments, experiences of harassment can be exacerbated by features such as synchronous voice chat, heightened feelings of presence and embodiment, and avatar movements that can feel like violations of personal space (such as simulated touching or grabbing). Simultaneously, efforts to govern these developing spaces are made more complex by the distributed landscape of virtual reality applications and the dynamic nature of local community norms. To better understand this nascent social and psychological environment, we interviewed VR users (n=25) about their experiences with harassment, abuse, and discomfort in social VR. We find that users' definitions of what constitutes online harassment are subjective and highly personal, which poses significant challenges for the enforcement of platform- or application-level policies. We also find that embodiment and presence in VR spaces make harassment feel more intense, while ephemerality and non-standardized application controls make it difficult to escape or report unwanted behavior. Finally, we find that shared norms for appropriate behavior in social VR are still emergent, and that users distinguish between newcomers who unknowingly violate expectations for appropriate behavior and those users who aim to cause intentional harm. We draw from social norms theory to help explain why norm formation is particularly challenging in virtual reality environments, and we discuss the implications of our findings for the top-down governance of online communities by platforms. We conclude by recommending alternative strategies for community governance.

CCS Concepts: • **Human-centered computing** → Human computer interaction (HCI); **Human-centered computing** → Collaborative and social computing; **Human-centered computing** → Virtual reality

KEYWORDS

Online harassment; online communities; virtual reality; social VR; embodiment; presence; moderation.

ACM Reference format:

Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proc. ACM Hum.-Comput. Interact.*, Vol. 3, No. CSCW, Article 100 (November 2019), 25 pages. https://doi.org/10.1145/3359202

Authors' addresses: Lindsay Blackwell (lblackw@fb.com), Oculus VR, Menlo Park, California, United States. Nicole Ellison (nicolee@fb.com), Oculus VR, Menlo Park, California, United States. Natasha Elliott-Deflo (natasha.elliott@oculus.com), Oculus VR, Menlo Park, California, United States. Raz Schwartz (raz@oculus.com), Oculus VR, Menlo Park, California, United States.

2573-0142/2019/11 – ART100 $15.00

https://doi.org/10.1145/3359202

1 INTRODUCTION

Harassment and other forms of abuse are a persistent problem in online spaces where social interactions occur. People who experience online harassment frequently report disruptions to their offline lives [6], including emotional and physical distress, changes to their future technology use, and increased safety and privacy concerns [20]. Online harassment often requires physical and emotional labor from targets, who must spend time documenting and reporting abuse in order for platforms to intervene. Although recent online harassment research remains focused on social media sites (e.g., [5,6,13,20,57,58]), early evidence suggests that abusive behaviors are occurring in similar ways in virtual reality environments [40,41,50], where experiences of harassment can be exacerbated by heightened feelings of presence and embodiment [52,55] and the real-time nature of virtual reality environments. Because the specific affordances of virtual reality environments can facilitate abuse in novel and unexpected ways, we hope that contributing a better understanding of the harassment experiences and support needs of social VR users can help technologists, researchers, and policy-writers develop regulation strategies for abuse behavior that anticipate—rather than react to—emergent technologies.

In this paper, we explore users' harassment experiences in social virtual reality applications, including the affordances which differentiate VR abuse from that sustained on social media sites, with a focus on better understanding the nuances of community governance in a nascent and transient context where concrete expectations for appropriate behavior have yet to emerge. We consider the following research questions:

> **RQ1:** What are users' experiences of harassment in social virtual reality, and how do they compare to experiences of harassment on social media sites?
>
> **RQ2:** What specific affordances of virtual reality environments exacerbate or mitigate harassment experiences?
>
> **RQ3:** What are users' expectations for appropriate behavior in social virtual reality environments, and how are these expectations established and enforced?

After conducting semi-structured interviews with 25 users of social VR applications, we find that users' definitions of what constitutes online harassment are subjective and highly personal, making it difficult to govern social spaces at the platform- or application-level. Participants' specific experiences of harassment in social VR largely fell into three categories: *verbal harassment,* such as personal insults or hateful slurs; *physical harassment,* such as simulated touching or grabbing; and *environmental harassment,* such as displaying graphic content on a shared screen. Because VR offers few but salient identity cues, we find that the identity signals that are available—e.g., dialect or gender as evidenced by voice via the audio channel—make some people more likely to experience harassment than others. We also find that the embodiment and presence afforded by VR can make harassment feel more intense; however, a few participants felt that embodiment and presence could reduce the incidence of harassment experiences by increasing empathy for other users, a finding supported by other empirical work [2]. We find that harassment experiences are further complicated by the ephemeral nature of virtual spaces and the lack of standardized controls across independent applications, which make it difficult for users to report or even escape unwanted behavior.

Finally, we find that shared norms for appropriate behavior in social VR are still emergent, with experiences varying significantly between social VR applications, due in part to the novelty

of the technology and high barriers to entry, but also because of the distributed nature of available applications and the transience of virtual social spaces. Because expectations for appropriateness are still unclear, we find that both moderators and ordinary users are reluctant to assume malintent on the part of individual violators, and that people distinguish between newcomers who unknowingly violate rules and those users who aim to cause intentional harm. We conclude by reflecting on the specific implications of our work for creating and governing pro-social spaces, focused on three major themes: facilitating the development of consistent pro-social norms; scaffolding community-led governance; and adopting an empathetic model of responsive regulation, all in the service of facilitating bottom-up—rather than top-down—behavioral enforcement.

2 LITERATURE REVIEW

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users, including but not limited to name calling and insults (sometimes called *flaming*); the release of personally identifiable information, such as a home address or phone number (*doxing*); or the use of another person's name or likeness without their consent [20]. Research into the prevalence of harassing behaviors on social media sites such as Facebook and Twitter has found that 66% of adult internet users in the United States have seen someone be harassed online, with 41% of all users having personally experienced some form of online harassment [20].

2.1 Harassment in virtual reality

Scholarship exploring early virtual spaces found that even in cue-sparse environments such as MUDs, in which users navigate space and interact with other users exclusively through text commands, female-appearing characters were "often besieged with attention," experiencing unwanted sexual advances from men who used visible lists of currently-active users to identify and approach users with female-sounding character names [10]. In 2016—more than 20 years after the publication of Bruckman's [10] early exploration of text-based MUDs—a woman named Jordan Belamire blogged about her first experience in virtual reality, using a friend's HTC Vive to play a game called QuiVr [4]. In the post—subsequently covered by several major news outlets (e.g., [61])—Belamire detailed how, minutes into her first multiplayer game, a stranger approached and virtually rubbed her chest and groin despite repeated requests to stop. Indeed, in a survey of frequent users of HTC Vive, Oculus Rift, Playstation VR and Microsoft Windows Mixed Reality [40], 49% of female respondents described experiencing sexual harassment in VR, including being groped, stalked, or catcalled; being shown a lewd photograph; or hearing a sexually explicit comment. 36% of male respondents described experiencing the same.

2.1.1 Toxic gaming culture and its influence on VR. Many of these reports on VR's "harassment problem" [45] implicate the toxicity of gaming culture, and specifically the frequent verbal insults associated with competitive play. Gaming environments, like other social spaces, reflect systems of structural oppression such as racism, sexism, and ableism: Gray [25] revealed racialized tensions between women in gaming environments in the wake of Gamergate, a term which both describes the controversy surrounding a loosely-organized community of disillusioned gamers and, colloquially, also refers to a specific faction of gamers who engaged in targeted harassment of female game developers and journalists over the course of several years. Similarly, Condis [16] attributes toxicity in gaming to hypermasculinity and community identity: those players who ignore these insults, refusing to take the bait, "demonstrate a cool-headed rationality, a mastery

over the self that is traditionally associated with the performance of masculinity." Players who react to or reject these insults—however demeaning—are instead perceived as "overly earnest and emotional," traits traditionally associated with femininity and with weakness. This, Condis argues, incentivizes players to engage in provocative and competitive language, both to establish membership in the community and avoid becoming targets themselves.

2.1.2 Affordances of VR which may exacerbate abuse. With gaming as a top use-case for virtual reality technology, and with both industries largely dominated by male users, the presumption that gaming culture has influenced virtual reality environments is not improbable. But even beyond cultural norms, harassment in virtual reality—as in gaming—is compounded by synchronous audio, where users levying insults are doing so in real-time and audibly, rather than on social media sites, where insults are issued asynchronously and primarily via text. Online harassment via more common forms of online communication (e.g., direct messaging) is not only asynchronous, but persistent and archivable—thus offering users more opportunities to control and mitigate harm, such as designating a surrogate to filter messages, automatically deleting them before they are seen, or reading them in bulk at a later time. Such options are not available when harassment is levied via a live audio feed.

Further, in virtual reality, users are embodied in avatars that move when the player moves and interact with other players in three-dimensional spaces, enabling violations of personal space and corporeal presence that feel fundamentally different than interactions that occur in other online environments—an experience made potentially more salient by the unique sensation of *presence* [52], or the feeling of truly "being there." Steuer [53] defines presence as the experience of our physical environments—not "as they exist in the physical world, but the perception of those surroundings as mediated by both automatic and controlled mental processes." Lombard and Ditton [36] identify several factors that can result in an increased sense of presence in mediated environments, including social richness; realism (do objects, events, and people feel "real?"); transportation (feeling as though you've been transported to another place); and perceptual and psychological immersion. VR environments offer heightened presence through a combination of some or all of these factors—meaning both positive and negative experiences can feel more "real" than in other mediated spaces, such as social media sites.

Finally, the potential for abuse is especially high in social VR applications, such as AltspaceVR or VRChat, which focus on general social interaction between users rather than on a shared game or experience (one example of the latter category is Oculus Venues, where users socialize while watching a live concert or sporting event). Most social VR applications facilitate interactions primarily between strangers, both due to the synchronous nature of VR (your friends may not be online when you are) as well as its relatively low adoption (your friends may not have access to a VR headset). While select virtual reality applications have lower-fidelity versions specifically for desktop or mobile access, most VR requires specific, and often expensive, hardware. Applications like Facebook Spaces, where users only interact with their existing Facebook Friends (i.e., known others), are less prone to abuse. When thirteen women were introduced to social VR for the first time [41], none reported experiencing harassment in Facebook Spaces—but when the same group of women used AltspaceVR, where central social spaces like a virtual "campfire" facilitate interactions between large groups of users who typically do not know one another, one woman was hit with a virtual stick; another was followed by a male avatar; and two women were sexually harassed. While social VR environments are still experiencing relatively low levels of participation, the incidence of negative experiences like these is likely to rise as VR hardware becomes more accessible over time. As norms that are established in the nascent stages of a

community can be powerful for shaping the perceptions and practices of later users, it is particularly important to explore emerging norms before adoption increases.

2.3 Regulating online communities

As in physical spaces, all online communities are regulated, whether formally (e.g., through the enforcement of established rules) or informally (e.g., through social norms). The first wave of community regulation, emerging in the 1980s [44], involved establishing norms for pro-social behavior and sometimes assigning community members special privileges (e.g., admins and moderators) to enforce those norms, often with the support of moderation tools such as reporting, flagging, and editorial rights [18,31,32,34]. A second regulatory wave introduced crowdsourced approaches, such as the decentralized approaches used by Slashdot and Digg [33,43]. While community-driven moderation approaches have been effective in smaller online communities [42], the size and scope of many online interactions have now largely outgrown normative regulation. In an effort to scale moderation practices, a more recent wave of regulation uses natural language processing and machine learning techniques to generate classifiers for detecting abusive language [12,29,62].

The success of automated regulation in rich, real-time and largely text-free virtual reality environments remains to be seen. Instead, because VR still has relatively few users—and because individual social VR applications have small userbases—most communities will be governed informally by emergent social norms, with some communities generating formal rules and designating moderators to enforce them. Because formal moderation practices in social VR are still developing and lack standardization across applications, these spaces are ripe for what Blackwell, Chen, Schoenebeck and Lampe [5] identify as a fourth wave of regulation, enacted when platform-level moderation fails: everyday users engaging in online harassment as a controversial form of social sanctioning. To mitigate that risk, our work examines current normative practices within social VR and their implications for the creation and successful governance of pro-social spaces.

2.3.1 Norms and social control. Social norms are the unwritten codes of conduct that influence behavior at both the individual and societal level [14]. Social norms differ from codified laws in that they are socially negotiated and learned through social interactions. Social norms—such as values, customs, stereotypes, and conventions—are "social frames of reference" that individuals first encounter through their interactions with others, and which later become internalized [49]. Perceived norms affect individuals' behaviors; collective norms affect behavior at a societal level.

Although the effects of social norms are widely studied, less is known about how and why norms emerge; however, the widely accepted instrumental theory posits that "norms tend to emerge to satisfy demands to mitigate negative externalities or to promote positive ones" [27]. Thus, norms are most likely to emerge when they favorably impact a given community's goals [39]. Suler [54] refers to the notion of cultural relativity: given the immense variety of online communities, "what is considered asocial behavior in one group may be very à propos in another." Traditional theories of behavior considered learning to be an individual process, governed primarily by reinforcement and punishment; social learning theory, however, proposed that learning can occur through observation or direct instruction [3].

Erikson [22] argues that communities use social norms to establish community boundaries—or rather, that those who misbehave establish community norms, which in turn influences how rules are made, enforced, and broken. Communities develop norms for appropriateness and enforce those norms through both formal sanctions, including formal policies and laws, and informal sanctions (such as shame, ridicule, disapproval, or ostracism, which all facilitate the regulation of non-normative behavior). Garland [23] argues that informal forms of social control exercised

through everyday relationships and institutions—e.g., by families, neighbors, and communities in schools, workplaces, and other social institutions—can undermine the authority of formal law by creating an "everyday environment of norms and sanctions" that is more visible and available than systems of formal legal control.

In one example of this, the participants of an online text-based community, LambdaMOO, experienced an obvious violation (an experience the author frames as "a rape in cyberspace") as impetus to come together and collectively decide on a framework for articulating and enforcing shared norms [18]. This piece, written 25 years before the present work, foreshadows some of the kinds of behaviors also enacted in contemporary virtual contexts—with a key difference being that the designers of early online communities were typically community members themselves, and thus had the ability to impose technical restrictions on behavior (e.g., a "boot" command to eject misbehaving users), particularly when a specific crisis necessitated an agile intervention. In the environments we study, this is not the case—a point we return to in our discussion.

Social virtual reality is a context that gives CSCW scholars the opportunity to revisit and reconceptualize our understanding of social norms, both as they emerge in relatively new (but growing) spaces and as they manifest given VR's unique affordances for communication. Exploring violations of norms within these environments—for example, harassment—presents an opportunity both to better understand norm development and enforcement, but also to strengthen our understanding of online abuse and the technical, social, or political infrastructures needed to appropriately protect targets and effectively sanction violators. Further, by exploring the specific affordances of these virtual environments that may facilitate or exacerbate abuse, we can ultimately advance solutions that are rooted not in the idiosyncrasies of a specific platform but which can instead be applied proactively to future systems as they are developed. By contributing a better understanding of harassment experiences in social VR, this research helps explain the shifting and emerging boundaries for appropriateness in this relatively new realm of human social interaction and inform the development of more universal mitigation strategies for online abuse.

3 METHODS

Because social VR applications are still relatively new—but given the rich history of CSCW scholarship about online harassment and similar forms of abuse—we conducted semi-structured interviews with 25 social VR users (all living in the United States), leveraging findings from prior research to structure our interview protocol. We chose a qualitative approach to more deeply and holistically examine the rich experiences of these users, particularly given the emergent context of these applications.

3.1 Recruitment

We conducted semi-structured audio interviews with 25 social VR users living in the United States. Our corporate recruiting team issued a general recruitment message ("We'd like to learn more about your experiences in social VR") via email to Oculus users in the US who had used a social VR application (e.g., VTime, Altspace, VRChat, or Rec Room) at least once in the past 28 days.[2] The recruitment message invited users to complete a short survey to indicate their interest in participating in the study. The survey explicitly informed potential respondents that none of their responses would have any impact on the status of their accounts, and that any personally

[2] L28, or the number of days active of the past 28 days, is an internal company standard for evaluating monthly usage.

identifiable information (such as names, specific employment, and account identifiers) would be stored securely and accessible only by the research recruiting team and researchers. The survey asked users their age, gender, and location; which social VR apps they had used, if any; whether someone had ever said or done something in social VR that made them feel uncomfortable; and whether they had ever said or done something in social VR that had made someone else feel uncomfortable.

The first recruitment survey had a total of 119 respondents, 18 of which (15%) were removed from consideration due to potential conflicts of interest identified by their employment information (i.e., employees of direct competitors were excluded, due to the sensitive nature of information exchanged during interviews). Of the remaining 101 respondents, 18 (18%) were women and 83 (82%) were men; no respondents identified as non-binary. In an effort to achieve better gender representation, we sent a second recruitment message, resulting in an additional 398 respondents, 77 of which (19%) were removed from consideration due to potential conflicts of interest. Of the remaining 321 respondents, 32 (10%) were women and 289 (90%) of whom were men; again, no respondents identified as non-binary.

Of the 422 eligible respondents, 91 (22%) reported having an uncomfortable experience in social VR. We contacted 37 of those 91 respondents to participate in a follow-up interview, based on a combination of factors, including age (we only contacted users aged 18 or older); use of social VR applications (we only contacted users who reported using multiple applications in which social interactions occur primarily between strangers); and reported availability to interview. Participants were compensated for their time with a $125 Amazon gift card, consistent with industry standards.

This study was approved by Oculus's internal research review. Participants were aware that they were participating in a company study, and they were assured that anything they disclosed during their interviews would not affect their Oculus accounts in any way, both during participant recruitment and again during the interviews themselves. While the company's communications team approved the resulting paper for publication, the researchers retained full control over the reporting of results.

3.2 Participants

We conducted interviews with 25 of the 37 participants we initially contacted, at which point we stopped hearing new themes and discontinued recruitment, using saturation as a criterion for halting data collection.

Twenty-three of our participants were men; only two participants were women, a limitation of the current work. Women are currently underrepresented in the total population of virtual reality users, which may be due, in part, to pervasive sexism in the technology sector [6,15,63] and because of the physical design of VR headsets may disproportionately cause motion sickness and other discomfort in women [37]. As women are more likely to experience the most severe forms of online harassment [19], future work should explore harassment experiences specific to women and non-binary VR users.

Four participants identified their age as between 18-24; six participants were between 25-34; nine participants were between 35-44; five participants were between 45-54; and one participant was 65 or older. Twelve participants were from the Western United States (Arizona, California, Oregon, and Washington); six participants were from the Northeast (New Jersey, New York, and Pennsylvania); five participants were from the South (North Carolina and Texas); and two participants were from the Midwest (Iowa and Minnesota). Participants held a wide range of occupations: five participants were students (two students also held full- or part-time jobs); five

participants worked in Arts and Entertainment; three in Information Technology; three in Engineering; two in Finance; and two in retail. Two participants were unemployed at the time of their interviews. The remaining participants represent a variety of sectors, including Private Security; Public Administration; Education; Health Care; and the Military.

3.3 Interviews

Interviews were conducted in August 2018 via phone by the first and second authors. Interviews lasted an average of 72 minutes, with the longest interview lasting 110 minutes and the shortest lasting 39 minutes. All participants were informed that their participation in the interview would remain confidential and would not impact the status of their VR accounts. Interviews were conducted by the first and second authors.

We first asked participants about their general internet and social media use, including participation in online communities (such as gaming communities). Then, we asked participants how long they had been using VR and about the specific social VR applications they used. Our participants used a wide range of social VR applications (see Table 2). The most frequently-used applications were AltspaceVR (n=21); Oculus Rooms (n=19); Rec Room (n=17); VRChat (n=13); Facebook Spaces (n=11); vTime (n=8); and Oculus Venues (n=7). For each social VR application, we asked participants why they used the application; how they found out about the application; what they liked and disliked about the application; and who they interacted with on the application. The protocol also included questions about participants' avatars.

We then asked participants about their personal experiences with social VR, soliciting specific examples of recent experiences participants had in social VR that they did and did not like. We asked participants about experiences in social VR that made them uncomfortable, including whether they felt they had ever been harassed or bullied in social VR; we also asked about what kinds of support (if any) they sought or remediation actions they took (e.g., blocking or reporting) as a result of their experiences. Participants were asked about witnessing and perpetrating uncomfortable experiences in social VR (e.g., "Have you ever done something in social VR that was viewed as inappropriate by someone else? How did you know that someone found it inappropriate?") and about their understanding of application- and platform-level rules. Finally, we asked participants about expectations for appropriate behavior in VR, including how those expectations are enforced, both formally (e.g., through moderation actions) and informally (e.g., through social sanctions).

Twenty-three of our 25 participants (92%) reported witnessing harassment in social VR (see Table 1). Eighteen of our 25 participants (72%) reported experiencing behaviors the authors would classify as harassment, including unwanted simulated touching or grabbing; violations of personal space; hate speech, such as racial slurs; personal insults; or displays of graphic or violent content on a shared screen. It is important to note that very few of our participants answered affirmatively when directly asked whether or not they had ever been harassed in social VR, despite also describing violating experiences like those outlined above. This is consistent with prior literature, which finds that people are unwilling to see themselves as victims; while people may hesitate to apply the label "harassment" to their own experiences, they readily apply the label to behaviors they witness others experiencing [6].

	Gender	*Age*	*Occupation*	*Location*	*Target*	*Witness*
P1	Man	35-44	Private Security; Student	CA	No	Yes
P2	Man	25-34	Engineering	NY	Yes	Yes
P3	Man	35-44	Information Technology	NY	Yes	Yes
P4	Man	18-24	Retail; Student	CA	No	Yes
P5	Woman	35-44	Arts and Entertainment	CA	No	No
P6	Man	45-54	Military	IA	Yes	Yes
P7	Man	≥65	Health Care	CA	Yes	Yes
P8	Man	35-44	Information Technology	CA	Yes	Yes
P9	Man	45-54	Arts and Entertainment	PA	Yes	Yes
P10	Man	18-24	Student	MN	No	Yes
P11	Man	25-34	Unemployed	TX	Yes	Yes
P12	Man	35-44	Engineering	NY	Yes	Yes
P13	Man	18-24	Student	CA	Yes	Yes
P14	Man	45-54	Retail	NJ	Yes	Yes
P15	Man	25-34	Student	CA	Yes	Yes
P16	Man	18-24	Arts and Entertainment	OR	No	No
P17	Woman	45-54	Unemployed	WA	Yes	Yes
P18	Man	35-44	Engineering	CA	Yes	Yes
P19	Man	35-44	Arts and Entertainment	LA	No	Yes
P20	Man	45-54	Education	NC	Yes	Yes
P21	Man	25-34	Public Administration	AZ	No	Yes
P22	Man	35-44	Information Technology	NC	Yes	Yes
P23	Man	35-44	Finance	NY	Yes	Yes
P24	Man	25-34	Arts and Entertainment	TX	Yes	Yes
P25	Man	25-34	Finance	NC	Yes	Yes

Table 1. Participant demographics.

3.4 Analysis

Interview recordings were transcribed by Rev.com and then imported into Dedoose. We employed an inductive analysis [56] to generate codes, generating an initial codebook based on recurring themes surfaced during interviews. After two members of the research team

independently coded one transcript to pilot the codebook, we iterated on our initial codebook, resulting in a total of 49 codes.

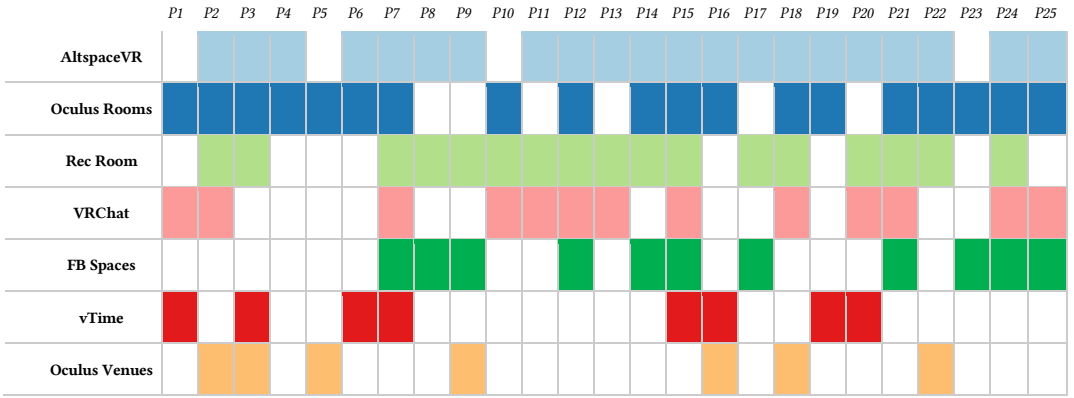

Table 2. Social VR applications used by our participants.

The resulting codes converged around several major themes, including defining harassment; specific affordances of VR (e.g., voice chat; presence; synchronicity); moderation and remediation actions (e.g., blocking and reporting; social sanctioning; moderator presence); expressions and manifestations of identity; and strategies for surfacing and enforcing expectations for appropriate behavior. Three authors independently coded transcripts using Dedoose, frequently discussing codes to maintain agreement. Quotations have been lightly edited for readability.

3.5 Limitations

As in other social spaces, systems of structural oppression (such as sexism and racism) result in people who hold marginalized identities (such as women and people of color) experiencing disproportionate amounts of harassment online [6,19,20,35]. Unfortunately, despite efforts to over-sample women and non-binary VR users, these experiences are not adequately represented in our sample. We also did not collect information about each participant's race, ethnicity, or sexual orientation; though some participants mentioned these details over the course of their interviews, we cannot assert that these findings are representative of the users who are most susceptible to abuse. This work, like all qualitative work, is also subject to social desirability bias; given the sensitive nature of our research, we hoped to partially reduce this bias by conducting interviews over the phone. Finally, this work was conducted by a single company located in the US; future work should examine our findings both in the context of other VR headsets and with international virtual reality users.

4 RESULTS

We find that users' definitions of what constitutes online harassment are subjective and highly personal, making it difficult to govern social spaces at the platform- or application-level. We also find that embodiment and presence make harassment feel more intense, while the ephemeral nature of virtual reality environments limits users' ability to successfully report abusive behavior. Finally, we find that shared norms for appropriate behavior in social VR are still emergent, and

that users distinguish between newcomers who unknowingly violate rules and those users who aim to cause intentional harm.

4.1 Definitions of online harassment are subjective and highly personal

Participants' definitions of online harassment were highly personal, although many participants described a similar rubric for deciding whether something was harassment or not: a behavior could be considered harassment if the person doing it continues after being asked to stop. When asked how he defined online harassment, P22 said: "Anytime you make someone feel uncomfortable. If someone asks you to stop and you just keep going, that's harassment in my opinion." P1 agreed, citing behaviors ranging from question-asking to death threats:

> "If somebody is asking you questions you don't want to answer over and over, that's harassment. If somebody's cursing a lot, or threatening verbally, that's harassment. If somebody is actually saying they're gonna kill you, or they're gonna go to your house, that's harassment. Racism. Stereotyping is harassment."

Other participants distinguished between behaviors that were simply annoying and those that were truly abusive. P17 said: "Harassment has got to be a little more than a guy who keeps texting you, 'Hi, hi, hi. I love you.' That was just annoying. Harassment would be, like, vicious and sexual communication that you didn't want." P24 described his own mental model for harassment severity, which helps him determine when to block someone versus simply ignoring them: "Say, like, a zero would be talking about somebody's mom. A five would be—this is me getting closer to blocking somebody—a five would be like a bad troll, just somebody that just keeps repeating themselves to get somebody roused up. Then a seven or eight—an eight would be people just saying disgusting stuff. Usually about a seven, I'm ready to pull the trigger." Although this participant categorized "talking about somebody's mom" as a 0—structurally not harassment—it should be noted that how users experience comments such as these is very idiosyncratic. For instance, a player who just experienced the loss of their mother might experience this as very hurtful; other players might read it as merely an attempt at humor. These personal and unpredictable responses to what may or may not be intended or experienced as harassment make it difficult to create universal definitions of "harassment" that work across communities and individuals.

Participants' specific experiences of harassment in social VR largely fell into three categories, which we categorize as: *verbal harassment*, such as personal insults or hateful slurs; *physical harassment*, such as unwanted simulated touching or sexual gestures; and *environmental harassment*, such as displaying graphic content on a shared screen or throwing virtual objects (see Table 3). For example, P18 was concerned with the potential for avatar-based harassment in VR, given the extent to which movement-triggered actions can make VR experiences feel more "real" than online spaces that depend on typing and reading. He explained: "I think there is a physicality to it, especially with avatars and realistic movement, that requires an extra level of rules around physical interactions." He also described situations in which a user obstructs another user's game or task: "Comments, lewd behavior, physical actions that make people uncomfortable. Getting up into somebody's face. I don't know if it really falls under harassment, but if people come in and ruin whatever activity that you're doing. Like if I'm trying to have a game of ping pong or basketball with somebody in Rec Room and then another avatar comes up, grabs the paddle or the ball and decides to then run off with it or chuck it across the room.

	Examples	*Medium*
Verbal	Personal insults; hate speech; sexualized language	VoIP; private messages
Physical	Unwanted touching; standing too close; obstructing movement; visible sexual gestures	Avatar movement
Environmental	Displaying sexual or violent content; drawing sexual images; throwing objects	Shared screens; virtual objects

Table 3. Types of harassment in virtual reality, as described by our participants.

Participants acknowledged that their past experiences online affected their perceptions of what constituted harassment in social VR. For instance, some participants said that their experiences with gaming culture, in which insults were commonplace, made them feel desensitized when witnessing or experiencing similar behavior in VR. This finding also helps explain why many participants did not respond affirmatively when asked directly if they had experienced harassment in social VR, despite describing specific experiences that align with current understandings of online abuse. P2 said: "I'm so used to that kind of thing being the norm in online games that it doesn't really register much for me." Similarly, P25 said that his experiences of online environments have always included harassment, to the point where he has come to expect it even as platforms and technologies evolve:

> "I can't speak for people who are black, gay, or Jewish, but I can tell you from my perspective of being a minority… I've been online since I was nine, ten years old. And I've played games online pretty much since then, and I've always come across this kind of behavior. And as online communities have evolved, this aspect hasn't really changed. So over time I've just kind of gotten used to it."

4.1.1 VR offers few but salient identity cues, making some people more vulnerable to harassment than others. Many participants felt that certain types of people—namely, women, children, people of color, and people who didn't have typically American accents—were much more likely to be harassed in VR than others. Although social VR offers fewer identity cues than other online contexts (such as social media sites), the identity signals that are available—e.g., dialect or gender cues embedded in voice—are powerful. These are also difficult to mask or alter, unlike other online communication channels where individuals can easily change profile images or textual descriptions of self.

Perhaps because vocal characteristics are difficult to disguise, voice was mentioned frequently by participants as a trigger for harassment. P18 said: "Basically the only kind of identifying thing you can have from somebody once they're speaking is their voice. I think gender is probably the most obvious thing you can get from the voice." P14 described an incident where a teenager with a brain injury was harassed due to his voice:

> "Maybe they don't speak very well, or maybe they don't enunciate. There was one kid, he was 18, he had a brain injury. So he did speak a little odd, and people would make fun of him until they realized. His father would go, 'You know, he has a brain injury, that's why he speaks like that.' And people would be more respectful because he had an injury than if he was, I guess, born like that. The way he spoke made him kind of an outcast in VR."

Many participants felt that women were more susceptible to harassment in VR both because of vocal cues and avatar appearance. P11 said: "Even if they're not using a feminine voice, just the same sort of general online harassment that comes to regular women I think comes to people using women avatars, despite the fact that they may very well be men." P18 agreed, having witnessed female avatars being physically harassed in public lobbies: "Someone came up to somebody that was a female—or at least had a female voice—in kind of the shared room, a room of probably 20 people or so, and went right up to her avatar and pretended to [perform a sexual act on] her."

Unlike voice, users do have control over the appearance of their avatars, and some participants chose to modify their avatar appearance to avoid potential harassment. P11, who is black, described different experiences when using social VR applications with a black avatar versus a white one. When playing role-playing games (RPGs), P11 chose to use a black avatar to more closely reflect his actual appearance. However, when using social VR applications, P11 ultimately decided to use a white avatar specifically to avoid racist harassment:

> "Since I'm going to be playing with a bunch of Americans anyway—and I can choose to get treated like a black person or not get treated like a black person—I'm probably going to choose not to get treated like a black person."

In addition to voice, the other salient and difficult-to-mask identity signal communicated in VR spaces is user height, which for headsets with motion tracking is captured by the headset's position in the room relative to the sensors and then expressed via the avatar. For our participants, the height of one's avatar was seen as a signal of whether one was a child or adult, another potential catalyst for harassment. P20 said: "I had a student that went into Rec Room. She was a tall graduate student. She was being harassed by little kids in Rec Room; it really noticeably bothered her. They were bugging her about her height. It was really obvious who the kids were and who the adults were, and she was getting teased for being an adult." Another participant said he could tell when children were playing because their avatars appeared shorter.

4.2 Embodiment and presence intensify harassment experiences—but ephemerality limits potential recourse

Although verbal harassment was commonly mentioned by our participants, many participants focused specifically on what they described as forms of physical aggression—in which avatars engage in unwanted actions upon one another—which some participants felt could be intensified due to heightened feelings of embodiment and presence. P2 said: "When you're in VR, you still feel like you have a sense of your body and your placement—so a bunch of people crowding you can actually make you feel a little bit anxious, even though it's all on VR." Another participant (P16) described harassment in VR as markedly different and more intense than similar experiences in gaming environments, where players only have live chat:

> "In VR, everything is live. You and that person occupy the space at the exact same time—with live feedback. You have the ability to literally look that avatar in the eye. You know what I mean? I think literally being able to see the person right in front of you has a sense of presence. [You] can really hurt somebody's feelings."

Other participants disagreed, drawing a distinction between online activities and those that happen in offline or "real" life. As P17 explained, "You hear stories about people in virtual reality

being uncomfortably touched. I guess I just don't get upset by things very much. It wouldn't be a big deal to me. In real life, it would be a much bigger deal, had they actually really physically touched me in an unwanted sexual way." As artfully described by Dibbell [18], this is one of the more complex issues for players (as well as designers and researchers) to navigate when considering VR environments: although participants experience incidents as traumatic—akin to offline, physical acts—VR actions do not result in harm done to one's body as would a physical assault.

Some participants also described feeling verbal harassment more intensely via voice chat than they would when reading text on a social media site. P5 said:

> "I just think that a lot more information carries over in voices. You know? You can tell a lot more about what someone means. Or they can disguise more of what they mean. Voice is a lot richer then text. I personally would find it a lot more creepy and scary, and I'm not sure exactly why—it's the same information. I think it would feel more confrontational. Voice is one step closer to someone actually being in front of you."

P9 felt most affected by the real-time nature of synchronous voice chat. He said: "It's more immediate. You are in that environment, and you are forced to participate in it in real time—as opposed to a message board, or on Twitter or Facebook. Something might turn up on your timeline, but you have to go there; it's not a real-time stream. It's a little more passive when it's just done on your phone or on a computer than it is in VR." Another participant (P24) described harassment in VR as feeling more personal than harassment on social media sites because it feels like being in public: "It's just one step closer to making it personal because you feel like you're somewhere. You're not just behind the computer screen in your room. You're out there almost in a feeling of being in public."

4.2.1 Embodiment and presence may also increase empathy and accountability. Although harassment in VR might feel more traumatic than similar violations experienced on social media sites, some participants felt that embodiment and presence could reduce the incidence of harassment by increasing empathy for other users (a finding supported by academic research, e.g. [2]). P2 said: "VR feels a bit more real—kind of a barrier to someone who, maybe on voice chat or something like that, wouldn't hesitate to say something bad." P3 agreed: "I think it's going to be harder for harassment to become a problem in VR. Virtual reality is a more humanizing experience. You're not just interacting with a flat picture and typed-out words. You're dealing with a human-looking avatar with three-dimensional presence that's interacting with you in real time. I think it's harder for people to be unresponsive to that." P2 went on to say:

> "You don't just have a sense of the person through their voice—you have a sense of the person through their mannerisms, through the way that they move their body, through how you see the avatars' heads or hands move. When somebody is sad, or feeling depressed or something, they tend to look down a little bit more. That translates into VR in a way that doesn't translate into text or voice chat. You have more social cues to work from—you have more levels of interaction with another person. That intensifies the interaction in VR and makes it more important to feel comfortable and to feel safe."

One participant felt that this heightened level of interaction translated into greater personal accountability in VR than in other online spaces. P10 said: "It's more personable. I feel like that

causes people to act more appropriately, because there are some repercussions… it's almost like you're face-to-face. Compared to [other online spaces], I feel like there's less harassment, simply because you technically *are* there in person—even though you're not."

4.2.2 Ephemerality limits potential recourse. Although harassment can feel more intense in VR than on social media sites, participants also felt limited in their potential recourse given the ephemeral nature of these spaces. P10 felt uneasy using existing reporting features, as any specific behaviors wouldn't be recorded or otherwise documented. While users can take screenshots in most VR applications, some participants felt it would be difficult to remember to use such a feature while experiencing harassment or similar emotional distress. Other participants found workarounds: P13, who had installed third-party gameplay recording software to share clips of his VR sessions with friends, was able to report "evidence" of a harassment experience using the same software.

Participants also described the lack of standardized moderation controls across individual applications as a barrier to reporting or even escaping harassment, with many participants relying on simply removing their headset—thus sacrificing their own VR experience—as the quickest and most efficient escape from unwanted or abusive behavior. P22, who had experienced harassment while engaged in a group conversation, said that he might have reacted differently if he had been alone: "I think I would have probably run away. You know, the whole fight or flight thing. The neat thing about VR is you just pick another place—or you just take the headset off and drop it." P21, however, described this same transience as an advantage for escaping unwanted interactions, especially when compared to similar interactions on social media sites:

> "When you post or reply to something on Twitter, it's there for the entire world to see. In VR, you're interacting with just who's around you, so it's easier to walk away. If you're in a room you want to get out of, you can walk away. The only option you have on Twitter is blocking someone or leaving Twitter."

This transience also made it difficult for participants to assess who was responsible for specific behaviors. P2 was attending a virtual comedy show when another audience member started making racist comments. Because of the real-time nature of voice chat and the large, ambiguous crowd, P2 couldn't identify which specific user was responsible:

> "You couldn't tell who was making those comments, so it wasn't very easy to report them. You can kind of hear everyone around you, so it's hard to know, 'Oh, it's this person, I should mute them.' It was one person making a bunch of really obnoxious, awful comments, but I ended up just muting the whole audience."

Sometimes, seemingly ephemeral interactions became more permanent. Several participants mentioned the practice of surreptitiously recording interactions in social VR and later posting them to YouTube, for the consumption of larger, unknown and unintended audiences. Recounting his experiences with a user who posted deceptive edits of recorded VR interactions to YouTube, P14 said:

> "I think that's the biggest thing that made me feel uncomfortable—the fact that I never knew what he was doing behind his avatar. One day, he put himself in the campfire, and you didn't know if he was filming or not. His avatar was just sitting there all day. I didn't want to be the one to show up on his YouTube channel."

This discomfort highlights an important tension: while ephemerality may limit a user's potential recourse when experiencing harassment in VR, participants presume—and in some cases, explicitly value—the impermanence of their own interactions. The reportedly common practice of sharing deceptively-edited recordings of interactions in social VR also highlights the limitations of seemingly persistent but manipulable media, particularly for systems which require users to submit evidence of perceived violations in order to sanction potential violators.

4.3 Fractured communities result in unclear norms for appropriate behavior

Social VR applications are still relatively new, and in some ways each application functions as a unique community, with a different set of emergent rules and nascent cultural norms interacting with more idiosyncratic understandings of what is or is not hurtful behavior. While most participants described relying on "common sense" to determine the boundaries for appropriate behavior, some participants described ways in which their personal expectations differed—sometimes significantly—from other users.

4.3.1 Users rely on "common sense" to guide their behavior instead of rules and policies. Most participants were aware that explicit codes of conduct for individual applications existed, but they could not recall a specific rule. P1 was frank about why he hadn't read the policy statement of one application: "There were so many little rules they were stating in each paragraph. It went on and on. You would just scroll through the page, and the text was so small, and you would scroll through the page and you would hit "agree" again, and again, and again. It makes no sense to do it that way." P1 had remembered one specific rule—that you can't disclose your exact address when sharing where you live—but only because another user had specifically reminded him. P5 said it had never occurred to her to look at the rules, "because I doubted I would break any."

Instead, participants described relying on their own personal notions of appropriateness, with many participants describing the boundaries for appropriate behavior as "common sense" or using the same general guidelines they would use for "real life" interactions. P19 described the expectations in VR as "a basic guideline that's kind of innate within us all." Said P18: "I don't remember seeing any rules, but I think it would be the standard 'no profanity, no racism, no hate speech'… basically no harassment. The kind of standard …'be cool' stuff." P10 employed a similar personal standard:

> "I try to keep swear words to a minimum. If I'm playing a game like Echo Arena, occasionally one will slip out, but otherwise I try to act like I'm with my grandma or something. If I wouldn't say it around her, I won't say it around somebody else."

P12 said that the more time he spends in a specific virtual place, the more responsibility he feels for its success: "I try to apply, as best as possible, the same social norms that I would try to conduct in the real world. The more you go to a virtual place the more it becomes like a real place. The more people you know, the more responsibility you feel for maintaining cultural norms and community standards. So, as a result, there's almost no distinction between my behavior there than in the real world." Here P12 describes the formation of community norms, which users often "import" from other contexts—in this case, "the real world."

Still, other participants acknowledged that their own personal expectations may differ from others'. As P12 said: "I do think I am more lenient towards trolling behavior. When someone is just running around, making noises, that can be harmless and somewhat entertaining. So, I do realize that my standards are a little bit different." P16 described an interaction where a stranger had asked him about his pornography preferences. He said: "Maybe that wouldn't offend

somebody else. I just don't like talking about pornography within the first five minutes of meeting somebody new. I guess that's what happens when you interact with real people. Some people value different things." P20 described generational differences in perceived appropriateness: "It did reinforce to me that there are a lot of millennials occupying these newer technology spaces, and they look at both language and how they behave in those spaces differently than I do."

Beyond individual differences, this "common sense" heuristic also breaks down when specific audiences are unclear. P23 emphasized the potential for interactions held within VR to reach larger, unknown and unintended audiences: "Sometimes I will have my kids around, playing around, and I forget to put in the earphones, and people may say something bad. They think there's only guys around, I guess. You don't know who's around the other side of the house, so you know, you have to watch your language. You never know who's on the other end of that, whoever you're interacting with."

4.3.2 Users distinguish between naïve newcomers and those who intentionally cause harm. Some participants found the lack of concrete norms in social VR liberating, particularly when using applications with little platform-level oversight. Said P25: "VRChat is personally my favorite social app. Not necessarily because they're doing anything great, it's just, like, craziness. It's kind of like the Wild West. There's no regulation, there's no moderation. People are just kinda doing their own thing." The "wild west" metaphor—evoking a time in American history with few laws and little enforcement of those that did exist, generally used colloquially to signal both freedom and lawlessness—was echoed by others.

Participants who had been using social VR for longer described initially appreciating the lack of formal rules or guidelines, but eventually choosing to invest in a particular community's success by helping to establish pro-social norms. Said P12: "A year ago or more, I felt a certain kind of freedom that came from just going into a virtual space and not feeling any sort of responsibility or a need to adhere to cultural standards or social norms. I think maybe my experiences have made me think more about virtual spaces. I guess maybe, the more time I spend in social VR, the better virtual citizen I'm becoming." Other participants described feeling a similar responsibility to their preferred communities, taking on moderation duties or helping to create onboarding materials for newcomers.

As new and more accessible devices are released, existing social VR applications often see an influx of new users, who are not yet acclimated to the norms of the space—making committed "virtual citizens" even more valuable for creating and enforcing boundaries. Many participants made a distinction between new, naïve users who unintentionally violate norms and users who cause intentional harm. P15 said: "You should get a second chance or even a third chance. People can learn to behave themselves, I think." Similarly, P13 made exceptions for people who may have made a one-time mistake: "Some people may look for people to make fun of… but also sometimes, people just aren't in a good mood, and they'll say stuff you don't like." These distinctions suggest a desire for more nuanced moderation tools that are responsive to varying motivations for participating in harmful behavior, particularly in spaces where norms are still in flux.

4.3.3 Moderators help establish norms for appropriate behavior. Finally, participants emphasized the importance of dedicated community members—whether they be formal moderators or simply volunteers—in establishing pro-social norms in VR, especially as communities continue to fluctuate in size and membership. P20 reflected on the importance of 'seed users,' much like Oldenburg's [38] "regulars": "Small changes draw different users—and they are shaping what that community looks like. Someone who shows up and doesn't like that experience is going to leave.

The first users, the newcomers… that first core of the few thousands of users will drive the experience of what it becomes later."

Many participants believed that individual communities should be responsible for determining what is and is not appropriate behavior. P12 said: "Specific to the platforms that I spend a lot of time on, I think there's a community investment that makes me a little more sensitive to keeping things peaceful and keeping things welcoming. I want people to feel like they want to be there." When asked who should be the judge of what is appropriate or inappropriate behavior, P21 said: "I guess some general rules, like any other social media platform… but I think the community itself, I would hope, would come up with their own rules."

P12, who is a moderator in AltspaceVR, tries to communicate normative messaging directly to users whenever possible. He said: "I feel a responsibility to try to get that person's attention in any way I can without being disruptive to the event. So, if the app has a text messaging service, I might use that to let them know that they're behaving in a way that really isn't right for the event. A couple times I've even invited the person to come to a different space so I can just talk to them. That works really well, actually. In almost every case where I've done that, they didn't even understand [the rules]. They were kind of just dropped in to something not understanding. Taking the time to personally explain that to someone can go a long way." The unique nature of the VR experience meant that some users did not at first understand the social contours of each new application or experience, sometimes believing they were interacting with AI bots instead of actual humans (further complicated by some applications' reliance on robot-like, rather than humanoid, avatars). For instance, P14 said: "People will come in and go, 'Are you real? Are you real?' And it's like, yes, we're real. This isn't a game—this is an actual social experience."

P3 also appreciated the hands-on approach of Altspace mods: "If you don't really present it as rules, then you don't get people who are tempted to break them. In AltspaceVR, they had a number of admins and guides who were very friendly, kind of chatted with you a little, helped you out if you were a little confused about what you were doing. I think there's ways to encourage friendliness to make it seem more of a value and not a rule. I think that's kind of important, because people love breaking rules—but when something is valued, it's perceived very differently. It's perceived as desirable, as opposed to something I have to do."

5 DISCUSSION

This paper uses qualitative data to understand how users experience harassment in social VR environments, including the specific affordances that contribute to these experiences. Our participants' experiences ranged from verbal (e.g., levying racist slurs via synchronous audio chat) and physical (e.g., simulated touching or grabbing of other avatars) violations to abuses of the environment itself, such as displaying graphic or disturbing content on a shared screen or throwing virtual objects. Within each category, participants described a diverse range of behaviors, from those perceived as simply annoying and easily dismissed—such as persistent question-asking—to more egregious and severe violations of comfort and safety, including racist remarks and even death threats.

Users' subjective definitions for what constitutes online harassment in social VR make governing these spaces at the application- or platform-level challenging—a finding consistent with research on social media users' definitions and perceptions of online harassment [6,19,20,35]. Top-down governance of social VR is further complicated by the diverse landscape of individual applications, which are owned, developed, and operated by independent developers and made available to users through hardware-specific "stores"—for example, the Oculus Store, where users

can buy individual applications accessible across Oculus-branded hardware (e.g., the Rift, Go, or Quest). This presents a unique governance challenge: because individual applications are not owned by singular companies (with a few exceptions), platform-level regulation is sparse, and platform-level policies that do exist are either obscured or otherwise not understood by everyday users. Indeed, following a comprehensive examination of macro- (site-wide), meso- (community-specific), and micro-level (individual) norms on Reddit, Chandrasekharan et al. [11] argue that macro-level norms can indeed "help moderators of new and emerging communities shape their regulation policies during the community's formative stages"—but only if the presence of such site-wide norms is known. Even if VR developers were to explicitly work toward platform-level regulation, the diverse population of independent developers, gaming companies, and corporations behind these applications are likely to have conflicting values and priorities that render top-down enforcement logistically impossible beyond the policies associated with the use of VR hardware itself.

This complexity, coupled with the cost of virtual reality hardware—making prototypical sanctions such as account- or device-level bans much costlier for companies to enact, should they incorrectly intervene—presents heretofore unseen challenges for top-down manifestations of both platform governance and community moderation. In the discussion below, we outline the specific implications of our work for creating and governing pro-social spaces, focused on three major themes: facilitating the development of consistent norms; scaffolding community-led governance; and adopting an empathetic model of responsive regulation, all in the service of facilitating bottom-up—rather than top-down—behavioral enforcement.

5.1 Facilitating the development of consistent pro-social norms

Our results have important implications for our understanding of how norms emerge in nascent—and often transient—environments. Previous scholarship has contended that users of traditional online platforms, such as social media sites, must reckon with large and unknown audiences when determining expectations for appropriate behavior, impeding the development of consistent and reliable local norms. Interestingly and in contrast to having only an ambient awareness of others, social VR users typically enter each space with an explicitly-defined and visible audience; these applications are both synchronous (i.e., every visible avatar corresponds with a user who is currently online) and physically bounded, allowing an individual user to easily and immediately assess their surroundings and audience. Information about social context is also more readily available in social VR than on social media sites: voice chat provides highly-warranted [60] clues to the identities of other individuals in a given space, and synchronicity enables users to quickly assess a new space. In theory, these affordances should accelerate the development of descriptive norms—but as our results demonstrate, users struggle to articulate a common definition for what constitutes an abuse of these virtual spaces.

This is further complicated by the complex landscape of VR applications, which are created by independent developers and trafficked in largely transient patterns (i.e., users are frequently moving between spaces, even within a single application). Any one user could conceivably be expected to, over the course of a single session in VR, be familiar with the formal policies and rules for a dozen independent applications—in addition to the policies for the software associated with the VR hardware itself. Given what we know about the significant gap between platforms' policies and their actual users' understanding of how platform-level enforcement decisions are made [17,24], it is unreasonable to expect any individual user to read, understand, and apply ever-changing policies for the full suite of virtual reality applications they choose to use. In the absence

of top-down enforcement across all social VR applications, our participants instead rely on "common sense"—that is, norms they have learned and imported from other contexts. Although the importing of norms from one context to another is natural, this complicates behavioral regulation within social VR for two reasons: first, the significant overlap between gaming communities and virtual reality communities meant that many participants interpreted verbal harassment and general hostility as holdovers from the more competitive, focused social environment of online gaming.

The second and perhaps more interesting complication is that the regulation of new media is largely driven by existing, familiar metaphors [26,30]; for instance, understanding the Internet as an online "library" evokes a set of regulatory and policy assumptions that other metaphors (e.g., a store) would not. Virtual reality presents a somewhat significant departure from other types of social experiences available online; scholarship and science fiction alike continue to position VR as key to the future of human social interaction, a philosophy Blascovich and Bailenson [7] characterize as the "Dawn of the Virtual Revolution." Indeed, most of our participants expressed excitement about the experiential possibilities of social virtual reality technology, where actions and experiences were not limited to those inhabited by immutable bodies constrained by laws of time and space. Given the interaction between undesirable behaviors that are becoming increasingly familiar in other online social contexts—such as harassment—and the relatively novel affordances of virtual affordances, the regulation of behavior at this nascent stage will be shaped by the appropriateness of the metaphors we use to understand these technologies and the social interactions they support. Future research should examine how developers, regulators, and users alike are conceptualizing virtual reality spaces, as a mechanism for better understanding the norms and customs that users will import into social VR spaces—particularly as the norms that develop early in these applications' lifespans will significantly influence the social mechanisms that will dictate and regulate behavior as adoption increases.

5.2 Scaffolding community governance

Given the complications of top-down governance across VR applications and the transient nature of these social spaces, our results underscore the importance of relying on community-led governance to regulate behavior. Users (or even moderators) of social VR are not the primary architects of the systems they are using, and as such enjoy limited opportunities to exert technical control over their own experiences and the experiences of their communities at large. Instead, our participants manipulated available identity cues—e.g., choosing an avatar of a white man—in order to customize their experiences and better insulate themselves to potential abuse. In the absence of individualized or community-owned controls, social VR applications can employ other strategies to facilitate governance from the community itself, including building infrastructure to support volunteer moderators. Many of our participants appreciated applications where visible moderators roam common spaces and reach out to new users (e.g., AltspaceVR), which they felt helped establish expectations for behavior more concretely than applications without moderators or greeters (e.g., VRChat). This higher-touch support for newcomers is especially important for communities who receive a large influx of new users, who may unintentionally alter the norms of the space if they cannot discern existing expectations and norms.

Application developers should consider incentivizing existing users to engage with and onboard new community members, and, if possible, implementing visible moderators who can model appropriate behavior in large social lobbies. Seering, Ng, and Yao [48] draw from the social identity theory of leadership to describe why group members who are "most prototypical of group

norms" often emerge as leaders. In contract to theories that attribute leadership capabilities to specific personality traits or access to resources, social identity theory suggests that a group's most pro-social members emerge as leaders in three defining phases [48]:

1) Self-categorization creates a spectrum of prototypicality within the group, with certain members deemed to be more prototypical than others.

2) Second, per the social attraction hypothesis, more prototypical group members are liked more than less prototypical members, and are thus able to exercise influence over other group members because individuals are more likely to help and support people that they like.

3) Third, group members make an attribution error by overattributing a leader's position to their personality characteristics rather than their prototypicality, reinforcing the belief that the leader possesses a particular disposition that helped them achieve their status within the group [48].

Among our participants, those who self-identified as community leaders (either as moderators or more informally as educators) indeed perceived themselves as sharing the larger values of their respective groups. Seering et al.'s [48] theoretical framing, coupled with this empirical result, suggests that designers could directly influence the norms of individual communities and groups through design "nudges" that encourage prototypical group members to engage more directly with new or norm-violating users. This suggestion is compatible with other recent research, which found that users of Twitch (a video-streaming platform) imitated examples of behavior they witnessed—especially behaviors from users perceived as having authority over the group or being otherwise high-status [47].

Further, rather than relying on platform-level rules or community guidelines to govern behavior, many participants felt that individual communities and their members should be responsible for establishing and enforcing their desired norms. We encourage designers to consider developing community-driven moderation tools that allow groups to establish and enforce their own boundaries for appropriate behavior. For example, Reddit allows individual subreddit moderators to create and enforce their own rules, regardless of how specific or even frivolous they may be (the subreddit r/ggggg only allows submissions that fully consist of the letter "G", in text or images[3]). Reddit has few rules that govern behavior at the platform-level, instead relying on individual communities to determine their own rules and trusting volunteer moderators to enforce them accordingly. A unique example of community-driven moderation is League of Legends' Tribunal, introduced by Riot Games in May 2011 to facilitate peer moderation. Reported users were assigned to other users for review, with reviewers examining chat logs, game statistics, and report details to decide whether or not the reported user should be punished, and if so, deciding collectively what an appropriate punishment might be. Riot found that the community's verdict was aligned with staff moderator decisions 80% of the time; in the other 20% of cases, players were more lenient than Riot staff [46], suggesting that bottom-up governance is a more empathetic approach to the regulation of social behavior online.

[3]http://web.archive.org/web/20190828182040/http://www.reddit.com/r/ggggg/

5.3 Adopting an empathetic model of responsive regulation

Perhaps counterintuitively, we find that the lack of a shared understanding of social norms in social VR makes users reluctant to categorize certain activities as problematic in intent, even when they are experienced as annoying or hurtful. In typical platform-driven moderation systems, all violators are treated equally, with users who unintentionally violate rules receiving the same sanctions as users who are deliberately trying to cause harm. In contrast, community-driven moderation approaches will allow users to accommodate individual differences, enabling peer-driven sanctions that offer well-intentioned users the benefit of the doubt and opportunities to reform their behavior.

An alternative model of formal norm enforcement is that of responsive regulation [1,8], a theory which aims to regulate undesirable behavior by being responsive to the conduct of individual perpetrators. Our results suggest that users already intuitively imagine diverse and varying punishments—a concept referred to in criminal justice as *proportionality* [5,59]—depending both on the specific type of violation and the perceived intent of the violator. For example, upon a violator's first infraction, a governing body might impose a counseling requirement, giving first-time violators the opportunity to reform their behavior and adhere to the normative standard. Should a first-time violator re-offend, however, the sanction should escalate proportionally. In a responsive regulatory pyramid (see Figure 1), the least interventionist punishments are applied to virtuous (i.e., potentially redeemable) actors, with sanctions escalating in severity until reaching total incapacitation at the top of the pyramid.

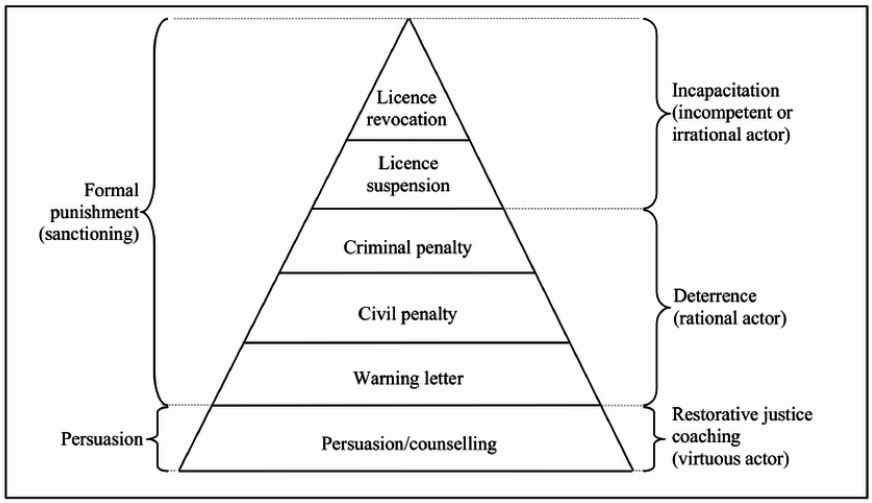

Figure 1. A responsive regulatory pyramid [28], adapted from [9].

A responsive system of regulation is not only a more empathetic approach to community moderation, but also helps applications to mitigate new abusive behaviors as they emerge. Social media platforms struggle to create and enforce additional policies to address emerging forms of harassment and abuse. For example, deepfakes—a sophisticated form of image manipulation used to fabricate video or audio "to show a person doing or saying something they did not do or say" [63]—has been documented by users as early as 2017, when a Reddit user used machine learning techniques to generate pornographic videos with the faces of Hollywood actresses. Deepfakes are often weaponized against women and girls as a form of "revenge porn," or the non-consensual sharing of intimate images [21]—but despite the proliferation of applications specifically designed

to easily generate explicit deepfakes, reliable technology for detecting when a video has been manipulated does not yet exist. As such, a user targeted by this form of harassment could be punished for violating platform policies forbidding sexual imagery at a time when they most need platform-level support. Using responsive regulation, platforms could apply less severe penalties to first-time offenders or when the full context surrounding a violation is unclear, allowing users who may themselves be targets to continue accessing critical social and structural support.

In virtual reality spaces, where communities are still growing and norms emerging, leveraging responsive regulation will empower communities to take swift action against violators without alienating users for incorrect enforcement decisions. Further, because responsive regulation gives violators opportunities to correct their behavior and adhere to the standards of a given community, platforms can feel confident taking harsh action (e.g., account removal or a device-level ban) against violators who escalate through every available sanction without the risk of rendering expensive VR hardware unusable for well-meaning users. Ultimately, community-driven regulation empowers communities to self-govern according to their own interpretations of appropriate behavior, fostering sustainable communities of users who feel a responsibility and commitment to the legitimate enforcement of consistent community norms.

5 CONCLUSION

Virtual reality environments present significant challenges for managing harassment and other abusive behaviors. We find that users' definitions of what constitutes online harassment are subjective and highly personal, and they include verbal attacks as well as violations of personal and physical space. Specifically, we contribute the notion of *environmental harassment*—or abuse committed through violations of the technical environment itself—to complement existing notions of both *verbal* and *physical* harassment. We also find that embodiment and presence make harassment feel more intense, while differing application controls make it difficult to escape or report unwanted behavior. This lack of a shared vocabulary or regulatory structure across individual social VR applications, when coupled with the novelty of the technology and the transience of virtual spaces, presents unique challenges for the development of consistent pro-social norms. We draw from social norms theory to help explain why the affordances of virtual reality make norm formation particularly challenging, and we discuss the implications of these findings for typical regulatory frameworks of top-down governance. Ultimately, we suggest bottom-up, community-led governance as one potential remedy, and we introduce Braithwaite's concept of *responsive regulation* to HCI literature as an alternative model of regulation which supports both community-led governance as well as more empathetic and strategic sanctions.

ACKNOWLEDGMENTS

We are grateful to Victoria Kyu and TJ Du for their contributions to this work. We thank our anonymous reviewers, whose suggestions helped improve this manuscript.

REFERENCES

[1] Ian Ayres and John Braithwaite. 1995. *Responsive regulation: Transcending the deregulation debate.* Oxford University Press, USA.

[2] Jeremy Bailenson. 2018. Experience on Demand: What Virtual Reality Is, how it Works, and what it Can Do. WW Norton & Company.

[3] Albert Bandura and Richard H. Walters. 1977. *Social learning theory.* Prentice-hall Englewood Cliffs, NJ.

[4] Jordan Belamire. 2016. My First Virtual Reality Groping. *Athena Talks.* Retrieved November 26, 2018 from https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee#.8lcy2o2bh

[5] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *Twelfth International AAAI Conference on Web and Social Media*.
[6] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW: 24:1–24:19. https://doi.org/10.1145/3134659
[7] Jim Blascovich and Jeremy Bailenson. 2011. Infinite Reality: Avatars, Eternal Life, New Worlds, and the Dawn of the Virtual Revolution. William Morrow & Co.
[8] John Braithwaite. 2002. *Restorative justice & responsive regulation*. Oxford University press on demand.
[9] John Braithwaite, Judith Healy, and Kathryn Dwan. 2005. The governance of health safety and quality. *Canberra: Commonwealth of Australia*.
[10] A. S. Bruckman. 1993. Gender Swapping on the Internet Proc. INET-93 ftp://media. mit. edu/pub/asb/papers/gender-swapping. txt.
[11] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 32.
[12] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.
[13] Gina Masullo Chen, Paromita Pain, Victoria Y. Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger. 2018. 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*: 1464884918768500.
[14] Adrienne Chung and Rajiv N. Rimal. 2016. Social norms: A review. *Review of Communication Research* 4: 1–28.
[15] Keith Collins. 2017. Tech is overwhelmingly white and male, and white men are just fine with that. Quartz.
[16] Megan Condis. 2018. Gaming Masculinity: Trolls, Fake Geeks, and the Gendered Battle for Online Culture. University of Iowa Press.
[17] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3: 410–428.
[18] Julian Dibbell. 1993. A Rape in Cyberspace. *Village Voice* XXXVIII, 51.
[19] Maeve Duggan. 2014. *Online harassment*. Pew Research Center.
[20] Maeve Duggan. 2017. *Online Harassment 2017*. Pew Research Center.
[21] Kirsti Melville for Earshot. 2019. "Humiliated, frightened and paranoid": The insidious rise of deepfake porn. *ABC News*.
[22] Kai Erikson. 1966. *Wayward puritans*. John Wiley & Sons, Incorporated.
[23] David Garland. 2004. Beyond the culture of control. *Critical review of international social and political philosophy* 7, 2: 160–189.
[24] Tarleton Gillespie. 2010. The politics of 'platforms.' *New media & society* 12, 3: 347–364.
[25] Kishonna L. Gray. 2016. Solidarity is for white women in gaming. Diversifying Barbie and Mortal Kombat: Intersectional perspectives and inclusive designs in gaming: 59–70.
[26] David J. Gunkel. 1998. The rule of metaphor: prolegomena to any future Internet regulation. *Electronic Journal of Communication* 8, 2.
[27] Michael Hechter and Karl-Dieter Opp. 2001. *Social norms*. Russell Sage Foundation.
[28] Norbert Hirschauer and Miroslava Bavorová. 2014. Advancing consumer protection through smart food safety regulation. *European Food and Feed Law Review*: 91–104.
[29] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]*.
[30] POOL Ithiel de Sola. 1983. *Technologies of freedom*. Harvard University Press.
[31] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA*.
[32] B. Kraut, N. Sherlis, J. Maning, T. Mudkophadhag, and S. Kiesler. 1996. HomeNet: A Field Trial of Residential Use of the Internet. In *Proceedings of CHI*, 77–90.
[33] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 543–550.
[34] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to Participate in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 1927–1936. https://doi.org/10.1145/1753326.1753616
[35] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute.
[36] Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: The concept of presence. *Journal of computer-mediated communication* 3, 2: JCMC321.
[37] Justin Munafo, Meg Diedrick, and Thomas A. Stoffregen. 2017. The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects. *Experimental Brain Research* 235, 3: 889–901. https://doi.org/10.1007/s00221-016-4846-7
[38] Ray Oldenburg. 1999. The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community. Da Capo Press.

[39] Karl-Dieter Opp. 2001. How do norms emerge? An outline of a theory. *Mind & Society* 2, 1: 101–128.
[40] Jessica Outlaw and Beth Duckles. 2018. Virtual Harassment: The Social Experience of 600+ Regular Virtual Reality (VR) Users. *The Extended Mind Blog* 4.
[41] Jessica Outlaw and Beth Duckles. Why Women Don't Like Social Virtual Reality: A Study of Safety, Usability, and Self-expression in Social VR. *The Extended Mind.*
[42] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, 51–60.
[43] Nathaniel Poor. 2005. Mechanisms of an online public sphere: The website Slashdot. *Journal of Computer-Mediated Communication* 10, 2: 00–00.
[44] Howard Rheingold. 1993. Multi-user dungeons and alternate identities. from The Virtual Community: homesteading on the electronic frontier, New York, HarperCollins.
[45] Katharine Schwab. 2018. VR Has A Harassment Problem. *Fast Company.*
[46] Dennis Scimeca. 2013. Using science to reform toxic player behavior in League of Legends. *Ars Technica.*
[47] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 111–125.
[48] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 1–34. https://doi.org/10.1145/3274771
[49] Muzafer Sherif. 1936. The psychology of social norms.
[50] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *Virtual Reality (VR), 2017 IEEE*, 225–226.
[51] Tom Simonite. 2019. Forget Politics. For Now, Deepfakes Are for Bullies. *Wired.*
[52] Mel Slater, Beau Lotto, Maria Marta Arnold, and María Victoria Sánchez-Vives. 2009. How we experience immersive virtual environments: the concept of presence and its measurement. *Anuario de Psicología, 2009, vol. 40, p. 193-210.*
[53] Jonathan Steuer. 1992. Defining virtual reality: Dimensions determining telepresence. *Journal of communication* 42, 4: 73–93.
[54] John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3: 321–326. https://doi.org/10.1089/1094931041291295
[55] J. Tham, A. H. Duin, L. Gee, N. Ernst, B. Abdelqader, and M. McGrath. 2018. Understanding Virtual Reality: Presence, Embodiment, and Professional Practice. *IEEE Transactions on Professional Communication* 61, 2: 178–195. https://doi.org/10.1109/TPC.2018.2804238
[56] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2: 237–246. https://doi.org/10.1177/1098214005283748
[57] George Veletsianos, Shandell Houlden, Jaigris Hodson, and Chandell Gosse. 2018. Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. *New Media & Society* 20, 12: 4689–4708. https://doi.org/10.1177/1461444818781324
[58] 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* (CSCW '17), 1231–1245. https://doi.org/10.1145/2998181.2998337
[59] Andrew Von Hirsch. 1992. Proportionality in the Philosophy of Punishment. *Crime and Justice* 16: 55–98.
[60] Joseph B. Walther and Malcolm R. Parks. 2002. Cues filtered out, cues filtered in. *Handbook of interpersonal communication* 3: 529–563.
[61] Julia Carrie Wong. 2016. Sexual harassment in virtual reality feels all too real – "it's creepy beyond creepy." *The Guardian.*
[62] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2: 1–7.
[63] Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. Retrieved April 4, 2019 from https://www.bls.gov/cps/cpsaat11.htmC

Received April 2019; revised June 2019; accepted August 2019.