Belonging There: VROOM-ing into the Uncanny Valley of XR Telepresence

BRENNAN JONES, University of Calgary, Canada
YAYING ZHANG, Microsoft Vancouver, Canada
PRISCILLA N. Y. WONG, University College London, United Kingdom
SEAN RINTEL, Microsoft Research Cambridge, United Kingdom

The world is entering a new normal of hybrid organisations, in which it will be common that some members are co-located and others are remote. Hybridity is rife with asymmetries that affect our sense of belonging in an organisational space. This paper reports a study of an XR Telepresence technology probe to explore how remote workers might present themselves and be perceived as an equal and unique embodied being in a workplace. VROOM (Virtual Robot Overlay for Online Meetings) augments a standard Mobile Robotic Telepresence experience by (1) adding a virtual avatar overlay of the remote person to the local space, viewable through a HoloLens worn by the local user, through which the remote user can gesture and express themselves, and (2) giving the remote user an immersive 360° view of the local space, captured by a 360° camera on the robot, which they can view through a VR headset. We ran a study to understand how pairs of participants (one local and one remote) collaborate using VROOM in a search and word-guessing game. Our findings illustrate that there is much potential for a system like VROOM to support dynamic collaborative activities in which embodiment, gesturing, mobility, spatial awareness, and non-verbal expressions are important. However, there are also challenges to be addressed, specifically around proprioception, the mixing of a physical robot body with a virtual human avatar, uncertainties of others' views and capabilities, fidelity of expressions, and the appearance of the avatar. We conclude with further design suggestions and recommendations for future work.

CCS Concepts: • **Human-centered computing → Mixed / augmented reality**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: telepresence, remote collaboration, video-mediated communication, hybrid meetings, mixed reality, augmented reality, virtual reality, remote embodiment, avatar, awareness

ACM Reference Format:
Brennan Jones, Yaying Zhang, Priscilla N. Y. Wong, and Sean Rintel. 2021. Belonging There: VROOM-ing into the Uncanny Valley of XR Telepresence. In *Proceedings of the ACM on Human-Computer Interaction (CSCW).* ACM, New York, NY, USA, 32 pages.

1 INTRODUCTION

Most research on the shortcomings of real-time 'hybrid' collaboration, in which some members are collocated and others are remote, tends to focus the glaring asymmetries of access in the moment. For a range of intersecting socio-technical reasons, those who are locally present are more likely to be included, while those who are remote are more likely to be marginalised [78, 79].

Much of this has to do with the disparity between remote and local participants' abilities to 'belong' to the space: to assert their presence in the space [79], and to make use of presence as a

CSCW 2021, November 03–07, 2021, Virtual event

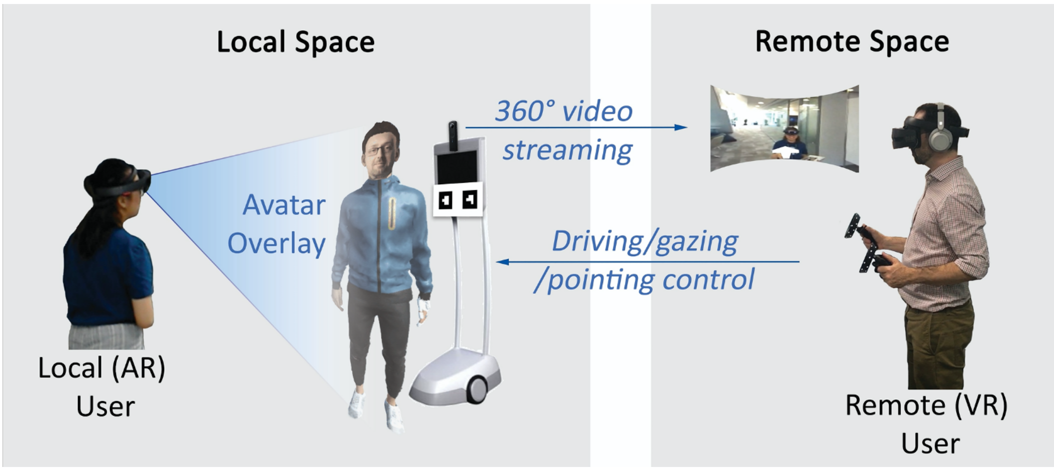

Fig. 1. An overview of the VROOM system.

collaboration resource, to deictically reference (e.g., point) [24, 28, 68], and to be aware of others' actions and presences (i.e., *workspace awareness* [32]). Remote participants need the same level of autonomy in a physical space as local participants in order to engage fully in certain types of meetings (such as brainstorming), and when the physical space is integral to the activity (such as a site manager's inspection of a work site for safety hazards or a choreographer's feedback on dance rehearsals in a theatre). Broadening out from a task focus,'belonging' is also characterised by social *affiliation* with one's colleagues, in the moment [60] and over time [14, 65], both directly and ambiently.

Remote workers fundamentally lack five of the building block affordances for physical autonomy and affiliation available to local workers: mobility, embodiment, expressiveness, awareness, and presence. A number of newer technologies exist to enable partial fulfillment of these needs – Mobile Robotic Presence (MRP), Virtual Reality (VR), Augmented Reality (AR), 360° video – but none of these fulfill all. In this paper we report on a study of an experimental Extended-Reality (XR) Telepresence technology probe [37] that combines these existing technologies in an attempt to fulfill these needs of autonomy and affiliation.

The probe, named VROOM (Virtual Robot Overlay for Online Meetings) [41], puts *belonging* at the centre of the experience. Instead of twinned endpoints, VROOM provides an experience contextualised to each endpoint (See Figure 1):

- For a person in the local activity space (Figure 1, left) wearing a HoloLens, an AR interface shows a life-size avatar of the remote user overlaid on a Beam telepresence robot using tracker markers. The Beam is also equipped with an extra 360° camera.
- For the remote user (Figure 1, right), a head-mounted VR interface presents an immersive 360° view of the local space. The remote user also has VR controllers, which allow for both piloting and gestural expression. The remote user has separate mobile and visual autonomy in the local activity space.

The remote user's speech, head pose, and hand movements, are applied to the avatar, along with canned blink, idle, and walking animations. The local user sees the entire avatar body in a third-person view. The remote user sees the avatar from the shoulders down in a first-person view. Local users can thus identify the remote user as a specific person, while the remote user has an identifiable embodiment of self, and both can take advantage of naturalistic verbal and gestural communication modalities. Together, these features enable a remote user to be recognised as an autonomous entity with which others can *affiliate*.

Masahiro Mori coined the term “Uncanny Valley" [59] to refer to the sharp dip in the continuous increase in affinity for entities with human likenesses. An industrial robot provokes limited affinity, a toy robot more affinity, but at some point, human likenesses such as human limb prosthetics provoke a strong sense of revulsion. AR and VR avatars can provoke similar reactions.

Given our emphasis on ‘belonging’ as comprised of autonomy and affinity, we argue that there is an ‘Uncanny Valley of Telepresence’ in which there a sharp dip in the continuous increase of both local and remote users’ senses of belonging as a telepresence experience becomes more like physical presence. The more the sense of belonging is provided, the higher the expectations, but also the more amplified the differences to physical presence become. Remote users expect more of their abilities, and local users expect remote users to have more abilities. Thus, when even simple problems occur (as they are bound to), people become more frustrated as their higher expectations are not met.

However, the Uncanny Valley is not a cliff. Mori included both the sharp dip and its equally steep recovery. On the recovery side of the function, Mori placed Banraku puppets just below a healthy human. Part of traditional Japanese theatre, Banraku puppets are less realistic than prosthetic limbs, but the gestalt of the theatre experience contextually enables audience members to focus on emotional performance. Thus, we argue that overcoming the ‘Uncanny Valley of Telepresence’ will involve finding the gestalt of what is most accountable about the comfort of belonging with others.

To explore these issues, we ran an exploratory study in which pairs of participants (one local, one remote) played two games (one involving gathering, the other guessing words) in two conditions: once using VROOM and once using standard Mobile Robotic Presence (MRP). Given that the MRP system was a mature commercial product and VROOM was a hack with significant limitations, the goal was not a simple evaluation of the success of one system compared to the other. Rather, the goal was to understand how users account for the systems’ different technological approaches to ‘belonging’ to a space, with special reference to finding the treacherous slopes and potential handholds in the ‘Uncanny Valley of Telepresence’.

Our findings illustrate that XR Telepresence has potential to support dynamic collaborative activities in which embodiment, gesturing, mobility, awareness, and non-verbal expressions in a physical location are important. However, there are also challenges to be addressed, specifically around proprioception, the mixing of a physical robot body with a virtual human avatar, uncertainties of one’s partner’s views and capabilities, fidelity of expressions, and the appearance of the avatar. Furthermore, the more immersive an XR Telepresence system is, the more amplified technical issues such as latency, video quality, and control become, largely because of the higher expectations of both remote and local users. We conclude that XR Telepresence should focus on providing the *comforts of belonging*, even if done in unnatural ways, rather than focusing on pure imitation. Aligning with Hollan and Stornetta’s [36] proposal that telepresence should seek ‘beyond being there’, we argue that the goal of XR Telepresence should be to help remote users *belong there*.

2 RELATED WORK

2.1 The Asymmetries of Telepresence

The asymmetries of telepresence are well established as wicked problems, especially in video-mediated communication, which represents the most common ‘high-fidelity’ version of telepresence in most workplaces [22, 34, 57]. In standard video-mediated communication, users are constrained in their abilities to achieve common ground [52], maintain awareness and control [88], and use spatial cues [82]. A remote person calling in to a meeting cannot control their viewpoint into the meeting room on their own, thus restricting them to seeing only what the laptop camera sees [27]. A remote user's lack of ability to control their view can lead to frustration and provide an unequal experience, making it difficult for the user to contribute to the activity at hand [40]. While there are some activities that can take place just fine with these asymmetries, other more-complicated activities suffer (e.g., where groups of people discuss and share ideas through sketching on whiteboards, refer to whiteboard drawings and objects in the room, and express ideas through body language). Hybrid contexts, as we noted above, tend to highlight these asymmetries [61, 79, 89]. VROOM aims to address five common issues of asymmetry: embodiment, expressiveness, mobility, awareness, and presence.

2.2 Embodiment

Embodiment is how the remote user is represented in the local space, how they perceive themselves and others, and how others perceive them. In a typical video call, the remote user's embodiment is simply a video image of themself on a computer screen. AR technology understands one's physical environment and overlays digital content on top of it. Applied in remote-connection technologies, AR can display the remote person in the local space, making the local person feel like they are together with the remote person.

A well-known example is Holoportation [67], in which a user can view the holographic imagery of another user through a Microsoft HoloLens head-mounted display (HMD), creating an impressively-convincing experience. For the holographic imagery, Holoportation uses depth and colour cameras to conduct realtime capture and representation of the remote user. Another approach projects the remote user into the space as a hologram [70]. Piumsomboon et al. [10, 71] and Rhee et al. [77] explored creating a 3D-modeled avatar of the remote user and mapping the remote user's movements to this avatar – an approach also taken by VROOM.

However, a common limitation of this kind of technology is the mobility of the remote person [45]. For example, when one user's room is laid out differently than the other user's room, conflicts can happen, for instance, one of the users may appear to 'walk through the wall' or 'stand on the table' in the other room. This is because the user is still physically navigating in their own space, and not in the other user's space. Further, users cannot explore a remote environment by themselves – they must be in a meeting with others. When they are, only others can see them, and only in the mapped mutual space. When the meeting ends, so too does the remote user's access.

A remote user's embodiment in the activity space affects not only how they are seen and perceived by others, but also their capabilities in the local space. Capabilities that a more-effective embodiment could enable include identity (e.g., others being able to identify the embodiment and associate it with the *person*) mobility in the space (ability to move around, control one's view in the space, or control where others perceive the user to 'be' in the space), expressiveness (e.g., through appearance, referencing/gesturing, speech, or other means of communication), establishment and assertion of one's 'space' or territory in the remote location (e.g., '*I'm standing here, you can't also be standing here*'), and physical capabilities such as the ability to manipulate objects.

Mobile Robotic Presence (MRP) systems, such as the *Beam* (by Suitable Technologies) [3], and the *Double* (by Double Robotics) [4], provide all five of these; though they provide limited identity, expressiveness, and physical capabilities. Research on the usage of telepresence robots in public spaces has found that people sometimes have a hard time identifying the person behind the embodiment, especially from afar and behind [54, 62, 63, 74, 85, 87]. Sometimes people identify the robot as a 'machine' or 'intruder', rather than a 'friendly person' [54, 90]. Similarly, while the mobility of these robots affords new ways for remote users to express themselves (e.g., through turning to 'gaze', 'gesture', or refer to something, or expressing emotions through 'dance' [90]), they still have a way to go before they reach, or perhaps even go beyond human levels of expressiveness. Lastly, telepresence robots provide few physical capabilities other than ‘rolling around’ on wheels. These limited capabilities have provided some frustration for users operating such robots in social settings and public spaces [35, 90].

2.3 Expressiveness

Expressiveness is the remote user’s ability to express themselves in the local space - to be seen and heard. The supposed value of video-mediated communication is that along with hearing the other, once can see their gestures, gaze, and emotional expressions such as facial expressions, posture, hand clapping, handshaking, and high fiving. People video chat often to show things to someone [64]. It is reasonably easy for people in the local space to show things to remote people, through either taking control of the camera and centering the object of interest in frame, or through placing the object in front of the camera [40, 55]. However, sometimes the remote user may want to show something in the local space to the local user, or draw attention toward something in the local space. With conventional systems this can only be done verbally, and quite often the remote user’s attempts to draw attention toward something in the local space are unsuccessful [40]. Researchers have proposed various ways of allowing remote users to gesture and refer to objects in the local space, including methods such as on-screen annotations [21, 23, 25, 26, 33, 50], showing hand gestures [49], telepointing [43], and AR avatar gesturing [71, 72, 77].

2.4 Mobility

Mobility is the remote user’s ability to ‘move around’ the local workspace, to control their view in it, or to at least explore different elements or things within it. A typical video-conferencing setup affords little to no mobility for the remote user, as they are restricted to the viewpoint of the camera on the device running the video-chat app. Some additional mobility might be provided if the remote user is able to ask someone in the room to pick up the device and move it around; although the remote user still needs to rely on a local partner and, as a result of their limited viewpoint, the remote user might not be able to identify opportunities to take control or find something of interest to get a better view of [40]. Local users could also put together a multi-camera setup in the meeting room, e.g., with one camera providing a good view of the attendees, while another camera provides a good view of the contents on a whiteboard. Telepresence robots are another solution. These have been around for some time, with early research on them dating back to 1998 [69]. They have been explored at workplaces such as Microsoft [80], as well as at conferences [63, 74], for museum tours [75], for connecting friends and long-distance couples [90–93], and in the outdoors [35]. While they are a step ahead of traditional video-conferencing interfaces for certain situations, especially those in which free exploration of a remote environment is essential or helpful, there are still limits to the amount of social and spatial presence they can provide. They still keep users essentially ‘trapped in rectangles’, not providing full immersion, and limiting the sense of social presence for the remote user (as well as limiting the local user’s sense that the remote user is ‘socially present’ with them).

2.5 Awareness

Awareness is the remote user’s knowledge and understanding of the local space, what and who is in it, and what everyone in the space is doing. Typical video-conferencing setups afford limited awareness, as the camera’s field of view (FOV) is limited, and the remote user is only able to see and hear what the device can pick up. While this is often sufficient for simple work meetings with fewer participants, it can be insufficient when there are more people, when the space is larger, or when the activity is more complex or dynamic, e.g., breakout-style activities, in which participants break into groups, or activities in which participants move around a lot and interact with numerous artifacts in the environment. Technologies such as 360° cameras for video conferencing (e.g., [35, 43, 86] help address challenges around visual awareness of the local space.

VR replaces a person's eyesight, hearing, touch etc. to create an illusion that the person is in another virtual space. Usually this virtual space is created via 3D modelling, like in fantasy scenes in VR games, but it can also be a real place captured by a 360° camera, such as for immersive 360° VR films. The 360° camera works as a "remote eye" for the user to see through and look around. This is also applied to remote-connection technologies, so that the remote person can see into the local person's place, providing them a feeling of 'being there'.

Moreover, instead of only letting the remote user view from a fixed location, researchers have explored attaching such 360° cameras onto the local user, so that the remote user can see from the local user's perspective [43, 86]. Additionally, researchers have proposed attaching a 360° camera to a telepresence robot, so that the remote user gains both immersion and autonomy to move at their own will [35].

2.6 Presence

Presence is the feeling of existing in a place, of 'being there'. *Telepresence* is the sense of existing in a place other than one's own [58], typically a real place, as compared to *virtual presence*, which is the feeling of being in a virtual, non-real, location [84]. Presence has many different facets, including *spatial presence* (the feeling of being *in* the space [53, 58]), *social presence* (the feeling of being *with* someone [15, 56]), *co-presence* (the feeling of being *in* a space *with* someone [38]), and *self-presence* (the feeling that one's 'self' or embodiment in the space is *indeed* themself [20, 29, 53]).

Traditional video-mediated communication provides limited telepresence, as it allows one to see another person and their space in a specific FOV, thus providing the abilities to see and hear someone in real time, and see into their space, their work, and some aspects of their day-to-day life. Other telepresence technologies take this further and attempt to bring one or more users 'through' this window, either to bring one user into the other's space, or to bring both users into a shared virtual or merged space. Telepresence can be enabled by many technologies, including telepresence robots (e.g., [3, 4]), drones (e.g., [39, 83]) VR (e.g., [35, 43, 77, 86]), AR (e.g., [71, 72, 77]), and holographic projections (e.g., [67]).

What ties this very large and complex body of work together is that it all points to ways of belonging. *Mobility* points to all users treating the local space as inhabitable and ownable by remote users who do not simply disappear at the press of a button. *Embodiment* points to all users treating the remote user as a human and not an object. *Expressiveness* points to all users being able to rely on a vastly expanded set of communicative resources to make meaning in the local space. *Awareness* points to all users being able to rely on the remote user's understanding of what is in the local space. Finally, *presence* points to the fundamental need for all users to feel that they are *there*, *in* a space, or *with* someone.

3 SYSTEM DESIGN

VROOM [41] (Figure 2) is a bi-directional asymmetrical XR Telepresence system. It is comprised of a telepresence robot (Beam Pro), 360° camera, AR headset (HoloLens), VR headset, and VR motion controllers. It is designed to enhance the mutual experience of telepresence for both a local (AR) and remote (VR) user. This system augments the experience of using a telepresence robot in two ways: (1) in the local space, by giving the local user (Figure 2, left) a view of a compelling representation of the remote user (Figure 2, right) and their gestures, head gazes, and other non-verbal behaviour, and (2) in the remote space, giving the remote user a more-immersive view into the activity space, allowing them to look around more freely in 360°. VROOM provides (1) through a virtual-avatar

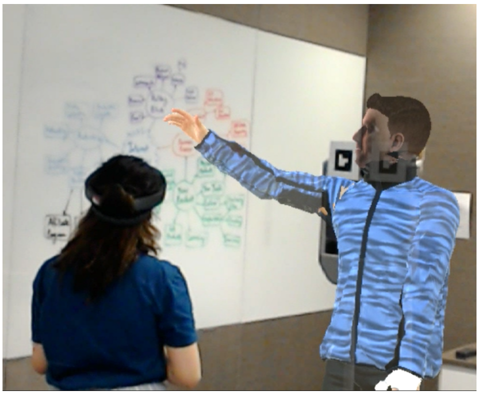

Fig. 2. A local (left) and remote (right) user using VROOM to collaborate on a whiteboard.

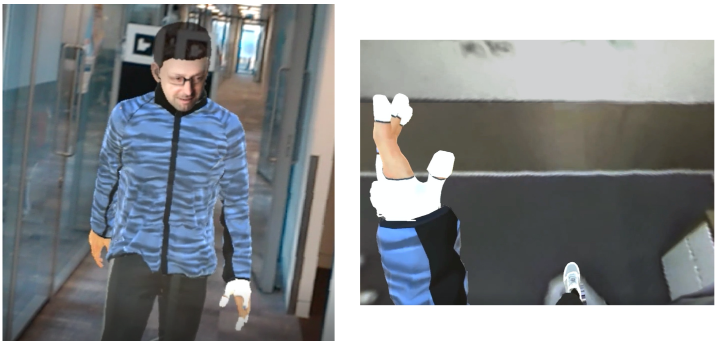

Fig. 3. Left: The remote user’s avatar overlaid on the robot, as it appears in the local user’s HoloLens. Right: The remote (VR) user’s 1st-person view of their own avatar, looking down at their avatar’s body.

representation of the remote user, viewed through a Microsoft HoloLens headset worn by the local user and overlaid over the telepresence robot (Figure 3 Left), and (2) through a 360° live video image, viewable through a HP Windows Mixed Reality VR headset, from a 360° camera attached to the top of the telepresence robot.

3.1 Local Space

The local space is the location where the activity is taking place. Here, the local user wears a HoloLens, through which they can see the remote user’s avatar superimposed over the telepresence robot.

The remote user is embodied in the local space through a combination of the telepresence robot and the virtual avatar overlay (Figure 3 Left). Marker tracking is used to detect the position of the robot and then superimpose the avatar over it in the HoloLens display. In the remote user’s VR view, the avatar is seen in first-person (e.g., when the user looks down, they can see their avatar’s shoulders, torso, legs, arms, etc., Figure 3 Right), and thus becomes like a first-person view of their own body.

The avatar’s appearance and actions are mapped to the remote user. A 2D image of the remote user’s face is used to make the 3D avatar’s head, so they have the same facial appearance (Figure 4). The avatar is rigged to respond to the remote user’s head and body actions, which are tracked

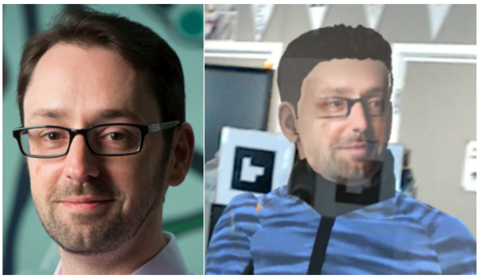

Fig. 4. The remote user's avatar's facial appearance (right) is made from a 2D photo of the remote user's face (left).

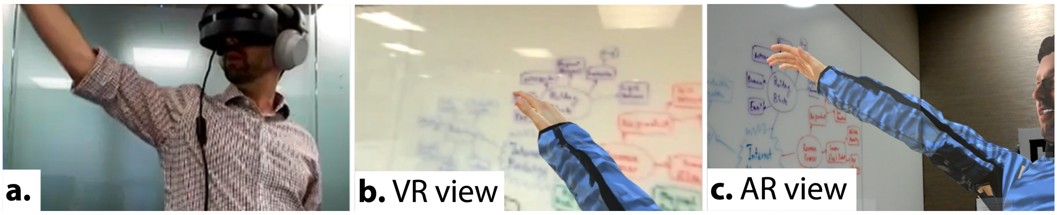

Fig. 5. The remote user gestures at a whiteboard. **(A)** The remote (VR) user's action. **(B)** The remote (VR) user's view. **(C)** The local (AR) user's view.

via sensors in the remote user's VR headset (head movement) and handheld motion controllers (arm movements). The head rotates as the remote user's head does, detected by motion sensors (gyroscope, accelerometer, etc.) in the VR headset. The hands and arms are rigged to move as the user moves Windows Mixed Reality handheld motion controllers. The remote user can press a button on each controller (left, right) to enable or disabled articulated arm movements for the corresponding arm. When arm movement is disabled for a particular arm, it is kept at the avatar's side. Driving the robot triggers a full body walk animation on the avatar, and an idle animation is applied when no locomotion or arm gesture input is detected. The avatar's mouth flaps in time with the remote user's speech. Lastly, a blink animation is applied periodically.

The avatar's movements and actions in both the local (AR) and remote (VR) views are synchronised, so where the remote user looks or points maps to where the avatar looks and points (Figure 5). This full-body avatar is meant to heighten the local user's sense that the remote user is present in the activity space with them. At the same time, the remote user's first-person view of their avatar body immersed in a 360° view of the space is intended to heighten the sense that the remote user is present with the local user in the activity space.

The robot has fiducial markers on the front and back for tracking [44] and a 360° camera attached to it to stream a live 360° view to the remote user.

3.2 Remote Space

The remote user wears a Windows Mixed Reality VR headset. This displays the 360° view from the perspective of the robot, as well as a first-person view of the avatar. The user also holds Windows Mixed Reality handheld motion controllers in each hand. One thumbstick drives the robot. Motion tracking on the handheld controllers animates the avatar's arm gestures. Since the remote user's hands are holding controllers, the arm movements are not very fine-grained nor high-fidelity, and there is no finger tracking nor finger animations. Remote users can activate or deactivate the hand tracking using the trigger on the controllers.

The remote user operates the robot using its built-in capabilities. On the UI of the standard Beam desktop app [7], the user sees a downward-facing view showing navigation guide lines indicating the projected direction of the robot's travel. In the VROOM VR view, we replaced these with a white arrow that was locked to the forward direction of the robot, to indicate which direction is forward for the robot in case the user loses orientation in VR.

3.3 Implementation

VROOM was built largely using existing technologies. In the local space, the user wears a Microsoft HoloLens (version 1) AR headset, running a custom-built application built with Unity that tracks the robot and superimposes the avatar over it. To track the robot, this app uses the HoloLensARToolKit library [13, 73] to track fiducial marker patterns [44] that were printed and placed on the robot. The avatar's head is made from an image of the user's face, using the Avatar Maker Pro Unity library [1], and attached to an animated human-body model available as a standard asset in Unity. The telepresence robot is a Beam Pro [3], which has a RICOH Theta V 360° camera [9] attached to its top. This camera connects to a small laptop attached to the robot's base, which runs another application that streams the 360° video from the camera to the VR application running on the remote side. On the remote side, the VR application is implemented with Unity and using an HP Windows Mixed Reality headset and handheld motion-controller set [8] connected to a Windows desktop PC. This application displays the 360° live video in the VR view, as well as a first-person view of the remote user's avatar. This app also sends the remote user's head orientation and arm-movement data to the local user's HoloLens app via HTTP polling. Lastly, the thumbstick on the remote user's motion controller sends driving commands (via mock keyboard commands) remotely to another PC in the remote user's space running the standard Beam desktop app [7], and the remote user wears a pair of headphones (with microphone) connected to the Beam app to speak through the robot and hear its surroundings.

4 STUDY METHOD

We conducted an exploratory study in which pairs of participants completed two rounds of the same game-like activities, differing in technological conditions of collaboration. For each pair, one participant was local and in the activity space, while the other was remote and operated a telepresence system in the activity space. Participants completed a pre-study demographic questionnaire and a post-study interview.

4.1 Participants

We looked for adult participants who were capable of either walking around an indoor open-office space or capable of using a VR and controller-based system. In line with other related qualitative and mixed-methods studies on video-conferencing and telepresence systems from previous work (e.g., [35, 39, 40, 48, 71, 86, 90]) we aimed to recruit around 8-16 pairs of participants, or about 16-32 participants in total, as this number is usually enough to find interesting interaction patterns and insights.

We recruited ten pairs of participants in the regional office of a global technology organisation, resulting in 20 participants in total. Participants were aged 25-54 (M = 35, SD = 7.3); with eight women and 12 men. Most participants (N = 13) identified as British, while two identified as American (from the United States), and one each identified as Chinese, Indian, Israeli, and Turkish. One participant preferred not to disclose their nationality. Most participants (N = 10) reported working in research. Of these ten, three were research interns, and one was a post-doctoral researcher. Five participants were software engineers, and five reported other occupations.

Despite being in a technology organisation, participants had varied experiences with the individual technologies of which VROOM was comprised, skewed to less experience. Four participants did not provide a response to these experience questions in the intake interview.

- Twelve participants reported using a head-mounted display (HMD) at least once per year. Of these 12, two reported using them weekly, and two monthly. Four participants reported having never used an HMD.
- Eleven participants reported using VR systems at least once per year. Of these 11, one reported using them on a weekly basis, and two on a monthly basis. Five participants reported having never used a VR system.
- Fourteen participants reported using game controllers at least once per year. Of these 14, one reported using them daily, three weekly, and two quarterly. Two participants reported having never used a game controller.
- Thirteen participants reported using a keyboard and mouse for playing video games at least once per year. Of these 13, one reported doing it daily, two weekly, three monthly, and one quarterly. Three participants reported having never used a keyboard or mouse for playing video games.
- Fourteen participants reported encountering 3D human-like avatars in video games and/or other apps (e.g., The Sims, Xbox avatars, etc.) at least once per year. Of those 14, six reported encountering them weekly, one monthly, and three quarterly. Two reported having never encountered them.

4.2 Study Design

We designed an exploratory study to compare and contrast how local and remote users experienced collaborating on an activity that involved moving around, searching, and being aware of the other's location in a physical office space, and a linked activity that involved focused conversation. Pairs completed the same activities in two conditions to compare a 'standard' MRP experience with the novel VROOM experience:

(1) The Beam condition utilised a Beam Mobile Robotic Presence (MRP) system, which is effectively traditional 2D video calling on wheels. The remote participant used the Beam's remote interface (the commercial app provided by the robot's manufacturer) on a desktop computer, which provided 2D video calling and piloting, controlled via a keyboard. The local participant used no personal technology, just interacted with the remote participant via the Beam robot's screen/speaker/microphone configuration.
(2) In the VROOM condition, the remote participant wore a VR HMD and headphones for communication and used the VROOM interface, seeing a 360° view of the activity space and a first-person view of their own avatar, and used two hand-held VR controllers for both locomotion control and moving the avatar's hands and arms. The local participant wore a HoloLens HMD to see the remote participant's full-body avatar overlaid over the Beam robot with the screen turned off.

As noted in our introduction, the Beam MRP system was clearly a mature technology compared to VROOM. This comparison was chosen because the Beam system was the closest alternative to VROOM, so it was hoped that participants would have less to learn overall and could also make very direct comparisons and contrasts.

4.3 Procedure

Pairs met in an office, and each participant first filled out a consent form and initial demographic survey. After this, the researchers explained the purpose of the study to the participants. Following this, they engaged in the *Beam* and *VROOM* study rounds, the order of which was counterbalanced across pairs. Local and remote roles were decided in the order participants signed up for the study and they did not switch roles during the study.

At the beginning of each round, the researchers explained the technology condition to both participants. For the *Beam* condition, this was an explanation of how the Beam robot worked, how to operate it, and what each user saw. Although robotic telepresence was novel to most participants, it was considered simple enough to understand without more demonstration.

Since the *VROOM* condition was very novel and also very different at each endpoint, more demonstration was provided for this. We explained to both participants how the remote VR interface worked, including the VR 360° view, the first-person view of the remote user's avatar, head turning, hand movements, and the ability to control the robot. We also explained how the local HoloLens interface would show the remote user's avatar. Remote participants were shown their avatar on the HoloLens to maximise embodied identification with their avatar. After being shown their avatar, remote participants were taken to a separate office and set up to use the VROOM system. The remote user was asked to look around in the 360° VR view, look down at the first-person view of their own avatar body, and move their hands. While doing this, the local participant was asked to observe the remote participant's head gaze and hand movements in order to understand how these movements were coordinated with the remote participant's actions. Remote users were asked to stand to begin with, but if they felt dizzy or nauseous they were provided with a chair with arms and could sit.

The participants then worked together on a game-like activity that consisted of two parts. In Part 1, *Gather*, they were asked to work together to find and take pictures of five items each in a relatively large open-office space. The local participant searched for five orange ping-pong balls, while the remote searched for five pink sheets of paper. They could talk to each other and help each other find these items. They were given 2½ minutes to find these items. Each item found added 30 seconds to the time they were given to complete Part 2 of the activity.

Part 2 of the activity involved playing *'Heads Up'*. Heads Up is a game in which Player A must get Player B to guess a specific word, without Player A saying say the word itself. The total amount of time we gave participants to complete this part of the activity depended on how many items they found in the previous part of the activity. Each item found provided 30 seconds of time, up to a maximum of five minutes. The local participant spent half of the total time as Player A while the remote was Player B. Participants switched roles halfway through their time.

Performance in Part 1 influenced the time pairs were given in Part 2 because we wanted to motivate the participants to perform at their best in Part 1.

In Part 1 of the activity, participants moved around the space, searched within it, gestured to things in the space and to each other, and looked to see where the other was located in the space. While the game itself might be unlike most work, the acts of which it was made up, or 'microactions' also make up facets of real work-based activities in which 'belonging to the space' is important. For example, moving around the space is important for activities that require exploration, such as workplace inspections and touring a new warehouse building. Positioning oneself within the room is important for conveying one's role and/or activity (e.g., 'I am positioned at the front of the room, therefore I am the leader and am about to speak'). Finally, gesturing and gazing toward objects are important for conveying intent or supplementing verbal communication via deictic referencing (e.g., saying 'that one' while pointing to a specific whiteboard in the room). The Gather game also provided a context for a similar range of microactions in both the Beam and VROOM conditions, but a clear difference in what might be possible in both conditions because the VROOM system theoretically had the advantage in terms of distant gestural capabilities. Part 2 enabled us to see how these experiences compared to a more conversational activity in which microactions in relation to the space are not as important or frequent, but expressivity is. Again, while word-guessing is not a common work activity, it was a structured activity which helped the participants interact and, in particular, focused one person's attention on the other. In this case, the Beam had the theoretical advantage of higher-fidelity transmission of facial and gestural expression.

After completing both the Beam and VROOM rounds, the participants were interviewed as a pair, in order to capture responses to one another's experiences. The interviews were approximately 30-minutes long and semi-structured. We asked participants a series of questions about their experiences working together in both the VROOM and standard Beam conditions, their perceptions of the avatar (both the remote participant's perceptions of 'self', and the local's perceptions of their partner), the remote participant's experience in the VR view, and any communication and collaboration strategies that participants used. The list of questions and prompts asked is in Appendix A. The exact questions asked to participants were based off this list, but in many cases we also asked follow-up and specific questions about events that we observed from participants during the study.

4.4 Analysis

The interviews were audio-recorded and transcribed. We used open, axial, and selective coding to analyse the interview data and reveal higher-level themes. Open codes included things like *local participants' perceptions of the remote's avatar*, *remote participants' perceptions of the VR view*, and *use of arm gestures*. Our axial codes included categorizations of the open codes, such as *awareness*, *expressivity*, and *proprioception*. From these, we landed on our selective codes, which include the higher-level themes of *embodiment*, *asymmetries*, and *mirroring* (described in the sections below). We used *video-based interaction analysis* [42] to analyse the video data. In the videos, we were interested in things such as the structures of events and participants' actions, turn-taking actions, the organisation of the activity, the spatial organisation of participants' interactions, and participants' conversations with each other during the study. One member of our research team analysed the interview data, while another analysed the video data. All four members of our research team (including the two who were primarily involved in the data analysis) met frequently to review the findings from the interview and video data and iteratively review and refine our codes. We also cross-referenced the interview codes with the key events from the video analysis in order to further iterate on our higher-level codes and assign our key events from the videos to our higher-level categorisations (our axial and selective codes from the interview analysis).

5 FINDINGS

Our findings are structured into four sections. First, we report the base representational needs for remote users to be seen as belonging to a local activity space: bodily identification and proprioception. We then report on two key communicative issues: bodily expression and communicative asymmetries.

5.1 Identification – Visual representation of remote users

Two of the most obvious challenges of telepresence are: getting remote users to identify with their visual representation, and getting others to identify remote users as unique individuals. In the Beam condition, remote users did not consider themselves as needing to identify with the robot, describing the experience as “detached" (P9) or referring to its nature as a machine. For example, P6 said “It was kind of just like a computer on wheels". By contrast, the VROOM condition was predicated on the remote user self-identifying with the avatar that overlaid the robot. After seeing what their avatar looked like in third-person (i.e., from the local’s perspective), some remote participants identified the avatar as at least somewhat resembling them in terms of its facial likeness and especially its arm motions. For some, as with Mori’s Bunraku puppets, the identifiably-personal *motions* of the arms were an important factor.

> *“It was my face, for sure."* - **P5-RU**

> *“[The] arms moved with my arms, and it looked where I looked, so yeah, I did identify with it. " – P4-RU*

However, other remote participants reported mismatches between their appearance and that of the avatar.

> *“Yeah, it looked kind of, I could recognise myself. [...] One thing was that like the torso, my breasts basically, seemed to be, this was the main thing I saw about myself, and it seemed to be much lower than seemed natural to me." – P3-RU*

> *“I don’t think I did identify with my avatar at all, really. [...] I just feel like it didn’t look like me that much. [...] When looking at it, I felt like it could have been anybody. I didn’t really feel like it was me, because it was the clothes and the hair and stuff." – P6-RU*

One reason for this mismatch is that the research team had to make the remote user’s avatar rather than participants customising it. The only changeable aspects of each avatar were a choice of just one male or female body, the semi-photorealistic face generated from a portrait photograph, and static hair that was manually matched from the portrait photograph. These limitations triggered an Uncanny Valley reaction when the avatar could be seen. Additionally, remote users only saw their avatar once and very briefly, rather than having a continuous or on-demand view, as mentioned by one participant.

> *“Yes. [Seeing my avatar through a mirror] definitely would have enhanced that sense of self. I don’t know whether it would have led to improved performance on the tasks, but as I said, I don’t think I was terribly aware of how I was being presented remotely. So interesting to think about ways that you could give that sense, yes." – P8-RU*

When it came to the local users identifying remote users, the Beam condition’s standard video-conferencing visual representation of the remote user made identification self-evident.

> *“The video gives you the facial recognition and the gestures, and so you could have a much more immersive conversation." – P7-LU*

> *“I looked at her a lot more when she was on the screen, and I could see it was actually [P6-RU], because the other thing was just a bit weird." – P6-LU*

As the last quote also shows, the VROOM condition’s avatar was treated as a clear deviation from the traditional video-conferencing view. The unnatural appearance of the avatar likely contributed to disconnection. P1-LU found “[the avatar did not] look exactly like [P1-RU]" and felt like they were “playing character[s]" in a “computer game". That being said, some local participants could at least identify the remote user’s VROOM avatar.

> *“It looked really close to him." – P7-LU*

> *“I could see that the face was trying to be a close match." – P6-LU*

But some local participants felt the avatar did not resemble the remote participant. This happened especially when the two participant knew each other well.

> *"So that avatar was better than I actually expected it to be, but it's still sufficiently different. Perhaps if we'd never met, if I'd never seen you [P11-RU], it might be again a different experience. But yeah, I was just aware it [the avatar] wasn't [P11-RU]" – P11-LU*

We also observed a disconnect between the remote user as a person and the remote user as an avatar. When P4-RU was trying to get P4-LU to guess the word "chair", both assumed that when P4-RU said that he was "*sitting* in one of these right now," he was referring to his actual body in the remote space, despite the avatar in the local space actually *standing* (see Figure 6).

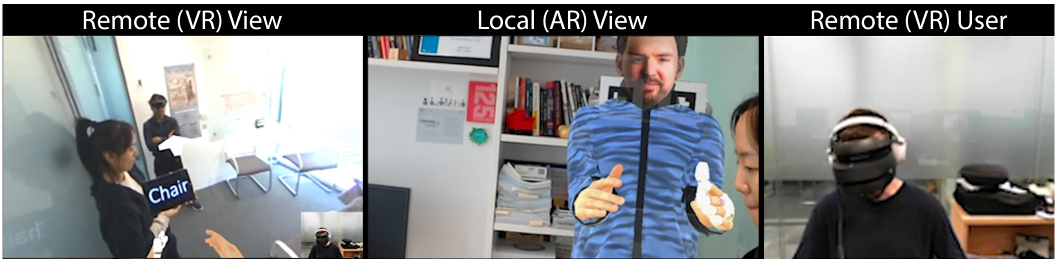

Fig. 6. The views of P4-RU avatar and P4-RU when guessing the word "Chair".

Later in the interviews, both participants said that they did not identify the avatar as in any way representing P4-RU.

> *P4-LU: [...] I might as well have done [...] that task with my eyes closed because I was just listening to what he was saying.*
> *P4-RU: So that's true, I wasn't aware, I knew that my avatar was standing but I didn't think about it at that moment, it was a disconnect.*
> *P4-LU: Yeah, and I was definitely very much thinking of [P4-RU] the person in the room 50 metres away, not [P4-RU] the avatar.*

Over half of the local users (P2, P3, P4, P6, P7, P8, P11 -LUs) reported that the FOV of the HoloLens was too narrow to enable a view of the VROOM avatar's face and body at the same time. This limited the value of attention on the avatar, and hence the building up of a strong sense of identification. Many participants (P1, P2, P6, P7, P8's LUs) also pointed out the lack of facial expression of the avatar. P6-LU thought that the lack of rich facial expression "take[s] away from the actual person" which made him "focus[ed] more on the voice cues than the visual." Indeed, half of the local participants (P3, P4, P5, P6, P7 -LUs) reported that they relied on mostly on voice while using VROOM. This is somewhat unsurprising given the long-standing primacy of audio in video-mediated communication [18], but here the lack of expression seemed to actively discourage visual engagement.

Despite these challenges, in line with previous studies that have found that a first-person avatar's appearance impacts the VR user's behaviour [46], the fact that there was a human avatar in the room appeared to have some subconscious impact on local users. P9-LU reported that in the VROOM condition, she felt more like she was engaging with another human.

> *"Even like the etiquette of how I treated [P9-RU], I feel like it's different. I noticed when [P9-RU] was an avatar person, I felt more like what I would do if she was physically next to me. Wait for her more and stuff like that."– P9-LU*

This effect was especially evident when orienting toward different directions. P2-LU pointed out that the existence of the avatar made it easier to tell left or right, relative to the remote user.

> *"I think when the avatar was revolving it was easier for me to say 'Something is to your right' or 'Something is to your left' than it was when it was the robot. [...] There was just less hesitation in saying it. [...] I think it's again because you can actually see something that is vaguely human shaped. Whereas before if it was just the robot you might think about 'Okay, how does that direction resolve in how they're piloting the robot'." – P2-LU*

The strength of this effect could actually be detrimental to the task if the HoloLens lost track of the fiducial markers on the robot, leading to a mismatch in the orientation of the avatar and robot. P11-LU tried to point out a target item to P11-RU when the avatar appeared to be facing the local user instead of aligning with the robot's orientation *perpendicular* to the local user (Figure 7). P11-LU provided navigation guidance to P11-RU relative to the *avatar's orientation*, directing P11-RU to turn to her left when the target item was actually to P11-RU's right in her remote view (see Figure 7).

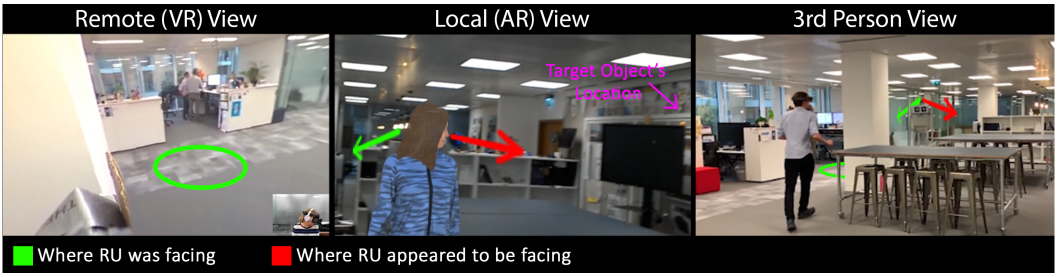

Fig. 7. P11-LU approaching and seeing P11-RU avatar facing him but P11-RU robot's back facing him.

5.2 Proprioception – Awareness of the Body in Another Space

While identification is important, a crucial step in ensuring that remote users feel that they *belong* in the local activity space is that they transfer their sense of proprioception (the awareness of the position of movement) of their real body to the avatar/robot body.

At one point, P2-RU was so immersed that when she wanted to move backwards she took a step back in her own space, rather than reversing the robot with the controller. This resulted in a disjunct experience when the robot did move with her.

> *"I think the good part is I think with the VR [head]set, I think I got more feeling of being in the room because it's harder to differ [whether] I go back or the avatar [goes] back with the controller. I physically moved myself to go back, this type of feeling. But at the same time, I do suffer a lot of motion sickness." – P2-RU*

Remote users reported that the 360° video increased their feeling of immersion (P1, P2, P3, P6, P7 -RUs), and suggested that they had "better vision" with VROOM's 360° view (P2, P4, P6, P7 -RUs). This confirms previous research showing that 360° cameras can increase immersion and awareness in video conferencing (e.g. [35, 43, 86]).

Remote users could also see their first-person avatar body moving with them in the immersive view. For example, P3-RU, although feeling that the avatar body was unnatural, still had a sense of proprioception because she could see her avatar's arms and legs moving as she tried to move. This confirms previous studies showing that synchronous visuomotor stimulation can create the illusion of body ownership [31, 47].

> *"So though at first I saw [my hands, I thought], 'Ah, they look weird and scary' but then I was 'Okay, they point when I want to point', so I quickly adjusted to it. [...] It was cool to*

> *see actually my legs walking from time to time, then I felt like 'Okay, I'm really there'" – P3-RU*

On the other hand, P6-RU reported that she had less of a sense of having limbs than eyes. For her, the VR vision was immersive enough to convey proprioception, but the illusion ended there.

> *"I imagined it [VROOM] more as like eyes, like it was eyes for me. It wasn't eyes and a body, so I could look around and see stuff. But I hadn't really connected that with, 'Well, I can use my body to point to a book that's behind me up on the shelf.' I don't know. I just didn't really put those things together." – P6-RU*

> *"Maybe I wasn't as aware as much, maybe I found it quite difficult to – I don't know if the word is embody – myself in the room. So I knew that there was a desk there, but I would never have pointed with my arm to that desk. I was still wanting to turn the robot and be like, 'that's the desk.' I felt like I found it difficult to know how I was orientated in the space, even though I saw my arms in front of me." – P6-RU*

Half of remote users reported that it was harder to navigate and manoeuvre in the VROOM condition (P3, P4, P5, P6, P8's RUs) than in the standard Beam condition. Difficulties stemmed from technical limits in the VROOM condition, such as using a thumbstick instead of a keyboard (P3, P4, P5's RUs), missing a dedicated floor-facing view (P3, P4, P5, P8, P11's RUs), and missing the Beam's trajectory guide (P3-RU). All of these things (the keyboard, floor-facing view, and trajectory guide) that the Beam has are unnatural, but despite the Beam lacking any first-person avatar appearance and 360° vision, some remote users felt that it provided a direct and expectable sense of proprioception compared to the VROOM condition.

> *"I talked about proprioception, the sense of where parts of your body are, [...] Because I was inhabiting this robot body, I needed to know where all the bits of the robot were. The robot interface was actually quite good at that, because it had a downward facing camera, so you could see the surroundings. Again, this is indirect and unnatural, but that gave me a pretty good sense of where I was as the robot relative to the rest of the room. In the VR setting, I was more tentative. [...] In video games or something, I think if you were presented as a different body, you could very quickly accept that as your own. But when that embodiment doesn't behave like you, doesn't have the capabilities that you have, and doesn't mirror your actions in the way that you would expect your body to work, I think that breaks the connection." – P8-RU*

Counter to our expectations, the unnaturalness in the Beam interface was actually more of an asset than the apparently natural view from the VROOM condition. The Beam robot's front camera view provided an obvious, if unnatural, *fish-eye view* plus a dedicated floor-facing view (Figure 8 left). The VROOM condition's immersive 360° view looked far more realistic, and users could turn their heads or look down (Figure 8 right).

The Beam's twin views seemed to provide an illusion of a larger sense of space coupled with the ability to flick one's eyes down to floor-facing view to check one's position relative to obstacles. VROOM, by contrast, required very active head movement because the naturalness of VROOM's view was an illusion. Although remote users could see 360° video, the VR headset only provided a 110° FOV compared to the human eye's 210° FOV.

> *"[I liked] being able to see the floor, and just having an eye on that while looking at the main thing was much better than having to – because I feel you have to have very exaggerated head movements with the VR [in VROOM] to be able to see." – P11-RU*

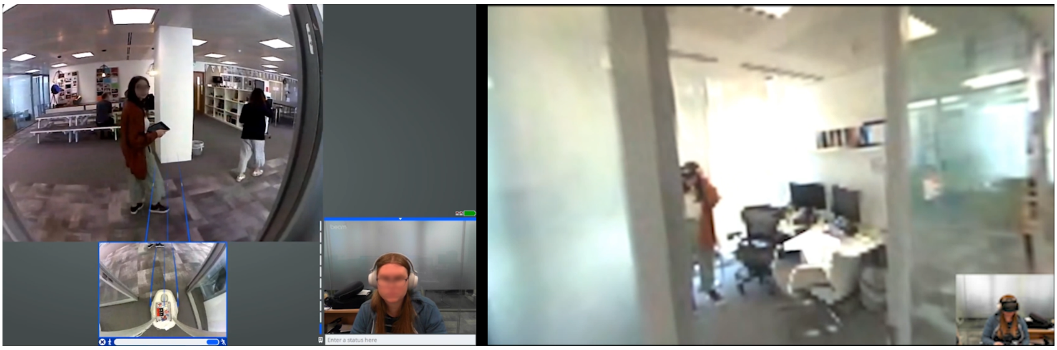

Fig. 8. P9-RU fitting through door with Beam (left) and VROOM (right).

> *“I struggled a bit with banging into things. Even though I knew I could turn my head and see my shoulders and I could look down and see my feet, I didn’t really do that, which was strange. So I felt a bit like I wasn’t sure if I could fit through things." – P5-RU*

The limitations of VROOM’s VR FOV also made it harder for remote users to see and use their hands. In real life, we can see our hands in our peripheral vision when we raise them at hip level. But in VROOM, remote users could only see their hands when they raised them chest high. This lead some remote users to disregard or forget about their hands.

> *“[I could see my hands] only when I brought them up here [to chest level], or if I looked down." – P7-RU*

> *“I felt like I couldn’t do enough with the hands to make them useful. [...] Because I felt like all I could [do] was lift them up and go like that, and then most of the time, [I] couldn’t really see them anyway. [...] I imagined it more as like eyes. [...] It wasn’t eyes and a body." – P6-RU*

5.3 Bodily Expression

The value in a remote user being able to identify with and have a proprioceptive sense of oneself is, of course, all in service of being able to communicate with people in the local activity space. The most relevant issue here is the ability to perform deictic and lexical gestures, collaborative gestures, and gestural mirroring.

5.3.1 Deictic Gestures. Deictic gestures are indicative or pointing movements [51]. As expected from prior research [35], deictic gestures at a distance were far easier in the VROOM condition than in the Beam condition. The VROOM avatar’s life-size 3D arms, while lacking hand articulation, enabled remote users to direct local users’ attention to things in the local activity space. P3-RU directed P3-LU’s attention to a pink sheet of paper with her right arm. In the later interview, P3-LU reported that he “really noticed" this instance and that he “could follow that [gesture] really well" (see Figure 9).

That being said, these deictic gestures could only be fairly crude and were not finely controllable.

> *“There was an equivalent with the waving in that I was getting what was happening because the arm was extending out of my field of view, and it was just flailing around rather than going like that, because the fidelity of the control." – P2-LU*

In the VROOM condition, the avatar’s hand tracking was turned off by default, requiring the remote user to toggle it on or off. We made this decision to give the remote user control over

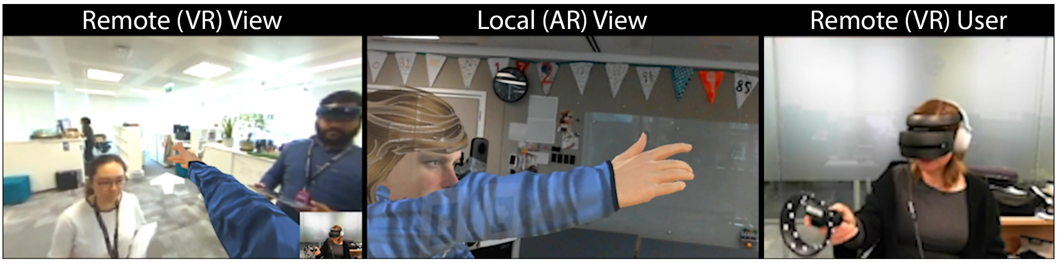

Fig. 9. P3-RU using deictic gesture to refer to an object location.

whether to show arm movement to the local user, and to limit unnatural arm movements. However, requiring the RU to actively initiate arm movement sometimes meant that remote users would makes arm gestures that were not communicated through the avatar because the remote user forgot to toggle the arms on. By comparison, the avatar's head movement was passively triggered. From participants' responses, we found that head orientation helped local users understand remote users' attention. For example, P3-LU felt that the avatar's head pose allowed him to "tell where the remote user was looking more closely." To some degree, this even remedied missing arm gestures, because people tend to look at something when they point to it. The P5 pair, who had previously used arm gestures well, experienced exactly that. When P5-LU asked the remote user whether she saw another pink target high up on the wall, P5-RU lifted her arm and asked: "There?", but she had not toggled arm movements on, so this gesture was not relayed to P5-LU. However, although P5-LU could not see the pointing gesture, she still recognised that P5-RU was attending to the pink target from the direction of her avatar's head gaze (Figure 10).

> *"When I was saying, 'can you see that?' I could see that you [P5-RU] were looking up, and I thought, 'she has seen that.' So that was quite helpful." – P5-LU*

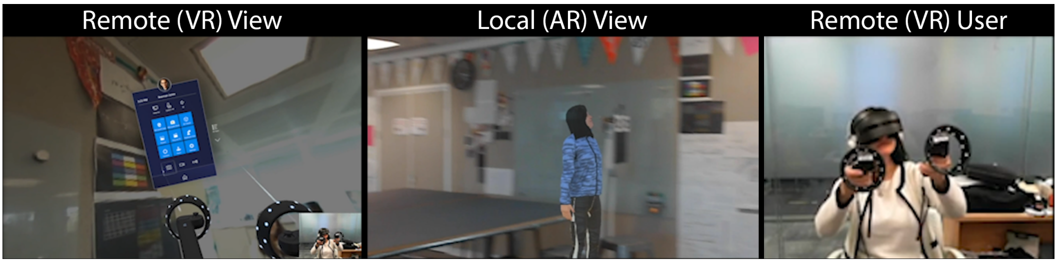

Fig. 10. P5-RU looking up to the wall.

One pair even managed to use head orientations to understand each other's intentions in action. P4-RU needed feedback on how far backward he should reverse the robot. First, he turned his head toward the study investigator, but received no response. He then turned his head toward P4-LU in front of him, and subsequently P4-LU responded.

While VROOM enabled larger deictic gestures, its lack of hand articulation was more noticeable when compared to the Beam condition at close quarters. Standard 2D video was able to show richer deictic gestures when people were facing each other, such as when P2-RU pointed at her lips with her index finger as she was describing the word 'lips' to P2-LU (see Figure 11).

Moreover, as we noted above, the communication of deictic gestures and head orientations in the VROOM condition was affected by the limited FOV of the HoloLens. Many participants (P2, P3,

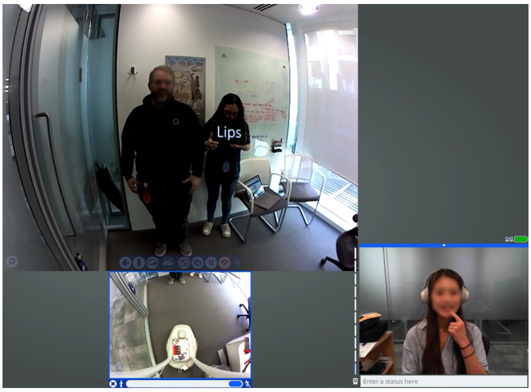

Fig. 11. P2-RU pointing to her lips as she's describing the word 'lips'.

P6, P7, P8 -LUs) reported that because their FOVs were small, they felt it was sometimes hard to see their partners' arm or head gestures. Some pairs even suggested that they decided to not rely on gestures in VROOM the condition because of the limited FOV (P6, P7, P8).

> *"I didn't get much from the avatar. I only saw the avatar when I was staring straight at it, and if I was staring straight at its head, I couldn't see his hands or anything else. So unless my face was totally focused right on the avatar, it was as if it wasn't there." – P7-LU*

5.3.2 Lexical Gestures. Lexical gestures are used to emphasise or elaborate on the content of one's speech [76]. Remote users in the VROOM condition attempted to use lexical gestures while holding their controllers. For example, P4-RU clapped his hands together vertically with his right hand on top of his left when describing the word 'staple' (Figure 12).

```
P4-RU: If I want to fasten two pieces of paper together, (positioning
right hand on top of left hand, clapping two hands together
vertically, see Figure 12) I would use this.
P4-RU: a paperclip. U:h staple.
P4-LU: U:h. (clapping two hands together vertically again). Yep.
```

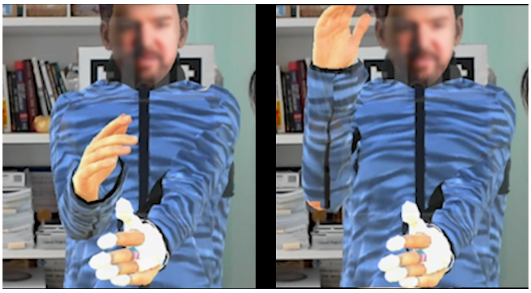

Fig. 12. P4-RU clapping his hands together vertically to describe the word 'staple'.

Lexical gestures were clearly more clumsy in the VROOM condition, and, unsurprisingly, at close quarters the Beam condition's simple video of the remote user was easier to understand.

5.3.3 Collaborative gestures and mirroring. Touch was of course impossible in both conditions, but in the Beam condition, participant pairs occasionally collaboratively mimed recognisable collaborative gestures such as high-fives (see Figure 13).

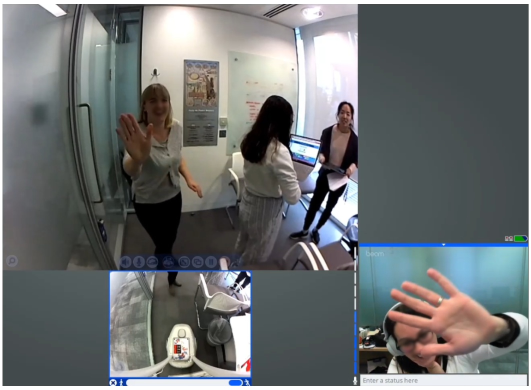

Fig. 13. P5-LU and P5-RU reciprocating high fives.

Perhaps more importantly than collaborative gestures, the ability to consciously or subconsciously mirror an interlocutor's gestures can be critical to a comfortable and engaging encounter. This, too, was difficult in both conditions. In the Beam condition, the constraint was largely the result of the limited FOV, with the remote user typically visible as just a head and shoulders unless deliberate gestures were in play. In the VROOM condition, gestural capability was limited to arms and head pose. The face was not articulated at all, and arm usage had to be very deliberate and might not even been toggled on. This lack of facial movement and hand usage was, ironically, reflected by local users.

> *"[Forgetting to use hands] could have been because I couldn't see your facial expressions, so I was totally neutral myself... I didn't use hand gestures at all to communicate with you, because it didn't occur to me, because I couldn't see your hand gestures." – P7-RU*

The main gestural mirroring that did take place in the VROOM condition was teasing remote users about awkward arm movements depicted by the avatars because of the way the controllers were seen by the VR headset. Such mismatches were possible when a physical body pose that was natural to the remote user was conveyed in an unnatural and distracting way to the local user. For example, when P5-RU drove VROOM in the hallway, she rested her arms on her desk. This posture was transferred to her VROOM avatar. The local user (P5-LU) found this amusing and briefly copied this pose (see Figure 14).

5.4 Communicative Asymmetries

VROOM is an asymmetric system in which remote and local users have different visual, auditory, and sensory experiences. This led to challenges in establishing a common understanding of each other's experiences, ultimately affecting how natural they felt with their own and the avatar's bodies, their interactions with another person, and the space.

The central challenge was that participants had limited awareness of what the system could show and what it was capable of. They also had limited time to explore these things on their own or tell each other about their experiences on their end. Without it being self-evident where the boundaries were, the quality of communication was affected. Local participants often could not see the avatars' full body expressions without putting more distance between themselves and the robot

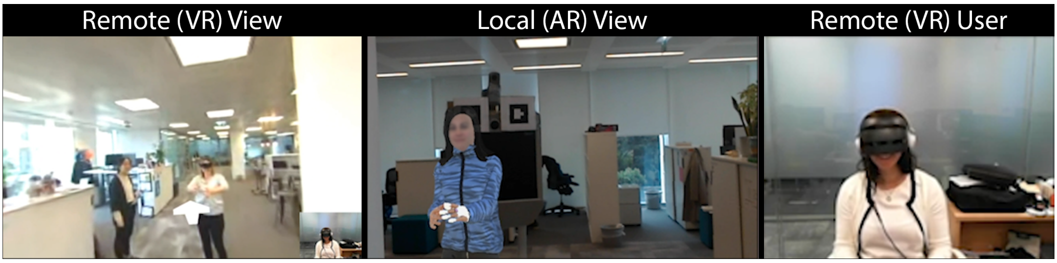

Fig. 14. P5-LU copying P5-RU's avatar's arms.

due to the limited FOV of the HoloLens. For instance, local participants would see the avatar's arms stretching out of their FOV when they were close to the robot. The interactional issue here is two-fold: remote users had to be aware of the need to adjust the positioning of their gestures to fit into the FOV while local participants needed to put more distance between themselves and the robot in order to see the non-verbal cues.

One remote user who had previous experience with HoloLens established understanding of her partner's limited FOV, and tried to adjust her behaviour accordingly.

> *"Yeah, it is very small. The face is very small. And if I was pointing with hands, I could see, [but] you might not have seen them. Yeah, that's a problem. [...] Because if he sees my face, he doesn't see my hands with the field of view." – P3-RU*

However, it was hard for inexperienced participants to understand the FOV of the HoloLens without experience. At least one remote participant explicitly chose not to use arm gestures on his avatar for this reason.

> *"For the reason that, the fact that I didn't really know what I was communicating with my hands because I didn't know what type of movements were actually being communicated, I didn't really use [hand gestures]. Especially in the word game where I could have used them, I didn't feel confident enough in them to actually use them." – P4-RU*

Similarly, while remote participants were unable to understand their local partners' technological experiences, local participants also often lacked knowledge of their remote partners' capabilities, including what they could see, and in what ways they were able to express themselves non-verbally.

> *"I was very unclear what he was able to see, so I didn't have a good visual understanding of how he was experiencing things. I wasn't sure, if I gestured, would he see it? And whether I was pointing, whether he would see it, those sorts of things." – P7-LU*

Many participants believed that if they had seen what the experience was like on the other side, they would have had greater understanding, awareness, and empathy for each other's experiences. Without the knowledge of the other person's experience, both local and remote participants reflected that they found communicating their capabilities and limitations difficult, especially when they were also supposed to work together to complete a task within a time limit.

> *"I would have liked a little bit more information like about what the experience was like from [P4-RU]'s side, because that was the biggest asymmetry I experienced. [It] was like I don't know what he can see, I don't know what controls he has, how much he can move around." – P4-LU*

> *"Yeah, I agree, it was hard to communicate what my limitations were. I mean obviously if I made a conscious effort, I could do it, but I was doing a task, right? So, it wasn't something I could do easily." – P4-RU*

One especially relevant asymmetry was that remote users often felt dependent on their local partners, as is common in MRP [17, 35, 90]. Remote users were unable to see or move around as easily as local users and unable to manipulate objects in the local activity space. This was the case for both conditions, but may have been exacerbated by expectations about VROOM's immersive capabilities.

> *"I feel there were two things going on. My visual signal was much reduced, and I had less sense of how I could move and where it was safe to move the robot. So I just felt more like a dependent participant than an equal partner." – P8-RU*

6 DISCUSSION

From our findings, we hypothesise that, just like Mori's Uncanny Valley [59], there is a similar notion of the 'Uncanny Valley of Telepresence': that the higher the simulation level of a telepresence experience, or the more it tries to simulate or replicate actual presence, the greater the user's feeling of 'belonging to the space', up to a turning point. At this point, the user's experience starts to deteriorate. The illusion of 'being there' gives them higher expectations of their own abilities in the space, but once those expectations are not fulfilled, they become frustrated and feel more as if they are not meant to be there. In other words, they feel less of a sense of 'belonging' there. Our study found that, while there are many potentials for an XR Telepresence system like VROOM, if such an uncanny valley exists, VROOM is probably somewhere within it.

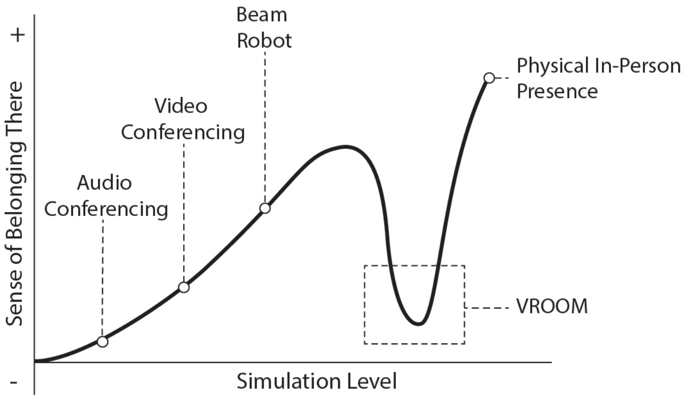

Fig. 15. The 'Uncanny Valley of Telepresence' hypothesis. While VROOM simulates in-person presence more than state-of-the-art tools like audio calling, video conferencing, and MRP, our findings suggest that it gave users less of a sense of 'belonging' to the space. Thus, VROOM is likely somewhere within the 'Uncanny Valley of Telepresence'.

Figure 15 illustrates our hypothesised 'Uncanny Valley of Telepresence'. Note that this is just for illustrative purposes and does not suggest precise measurements, nor are the axes to scale. The x-axis represents the degree to which real presence is simulated. At the end of the x-axis is physical in-person presence. The y-axis represents the degree of feeling like one 'belongs' to the space. VROOM has a higher simulation level than the standard Beam interface. However, the increased expressivity of VROOM likely led both remote and local users having higher expectations of the remote user's abilities in the space, though with most of these expectations not being met. Indeed, for some, the experience was effectively downgraded to the level of voice calls – e.g., P3, P4, P5, P6, and P7 -LUs relied on just voice more while using VROOM than while using the standard Beam interface. Similarly, in terms of mobility, remote users felt more dependent on their local partners while using VROOM than while using the standard Beam interface. This increased dependence, and thus decreased autonomy, again likely leading to a lesser sense of 'belonging' to the space.

There are three particular challenges with VROOM that we found contributed to this decreased autonomy: (1) limits to the embodiment's appearance, expressivity, and proprioception, (2) uncertainties about the other partner's capabilities, and (3) 'typical' technical challenges in video-mediated communication, which are further amplified when the experience is immersive.

6.1 Limits to the Embodiment

Remote users recognised themselves as being embodied by the avatars to some extent, but they did not necessarily recognise their embodiments as fully *belonging to them*. One reason for this was the appearance of the avatar, which was an off-the-shelf model rather than a customised representation. The remote participants in our study did not make their own avatars, but rather, they were made from front-facing facial photos that they sent us before participating in the study. Allowing users to create their own avatar, and perhaps resorting to less-photorealistic avatars such as Xbox avatars [5] or Microsoft Rocketbox avatars [12, 30], could have allowed users to identify with their avatars more closely, as they may have felt more attached to an avatar that they spent a lot of time creating. This might lead to avatars that are less realistic copies of their users, but if the point is to increase belonging, then pure imitation is not the threshold, but rather negotiated agreement over what works and is appropriate to represent the self is.

As well as not being very representative of remote users' actual bodies, the first-person view was also not as visible to remote users as their real body would have been. For example, remote users had to hold up their hands to see them because the VR FOV, while it looked convincingly immersive, was actually less than their human-eye FOV. Similarly, looking down to see their torso, legs, and feet, or left and right to see their shoulders, had to be active choices. We modelled the first person view of the avatar as a twin of the third person view, with all the same proportions. We wonder, however, if the first person view should in fact be distorted through enlargement and other procedures to be more noticeable in the first person FOV - much like the player's view of self in first person video games. Despite this being an unnatural, non-imitative change, it might improve the sense of identification by enabling constant peripheral views of one's body in the local activity space. This might reduce the strength of the feeling that one was “just eyes", as P6-U said.

The VROOM avatar's lack of expressiveness was also frequently cited by both local and remote participants as a challenge. While value was found in the remote user being able to point and gesture at a distance in the space and the local being able to see and understand those gestures, participants thought that these gestures were not rich enough nor had high-enough fidelity. Richer expressions such as shoulders and torso moving when looking left or right, finer-grained finger movements, and richer facial expressions were missing. While some of these may be less important for functionally collaborating in or 'belonging' to a space, as they do not all involve referring to objects or locations in the space, these non-verbal expressions are certainly important for social cues and empathy, and thus for social presence [16]. Remote participants frequently stated that they were able to make many of these non-verbal cues in the standard Beam interface at close range, and that it would have been even better if they were able to make these richer expressions in VROOM, so that their partners could see them in 3D and from multiple perspectives. Technologies that could enable these already exist. For example, VR headsets such as Oculus Quest can track hand and finger poses [6], as can the HoloLens 2 [81]. Microsoft Azure Kinect depth cameras [2] can track hand and finger poses as well as full body poses. These could combine to holistically apply fine-grained hand and full-body movements to remote user's avatars. Facial expressions are difficult to capture when both remote and local users are wearing HMDs, but new approaches with cameras capturing upper and lower face expressions are being developed [66].

Finally, *proprioception* is key to the sense of belonging there. Remote users had less proprioception using VROOM than using the standard Beam interface. The Beam robot is not like a human in terms of either appearance or facial/bodily expressions, but the Beam robot's behaviour, as well as the standard Beam interface, are more controllable and predictable. When given an interface designed *for piloting the robot* – the two camera views providing bodily awareness of the robot's edges and corners, navigation guide lines providing awareness of where the robot body will end up, and a fish-eye view of the world in the top view providing a deeply unnatural but somehow comforting feeling that physical objects in the world were further from the edges of the robot – the remote users were left with no surprises, and they *expected* an unnatural yet predictable arms-length experience. Conversely, the VROOM interface immersed the remote user's vision in the local activity space and showed a first person virtual human avatar overlay, but remote users still physically controlled the robot in the local activity space. The visual illusion was convincing enough to disembody vision, but had no further illusions for movement.

Mori [59] hypothesised that the lack of animation of entities reminds us of the stillness of death. As an entity is more real, the nuances of its stillness-within-movement provide a feeling that whatever is animating the entity is unnatural. That is, the issue is not simply stillness or movement, but a problem with the accountability of that movement. However the *context* and *placement* gestalt of Banraku puppets in a theatre experience allowed the human-*driven* movements of Banraku puppets be accounted for as emotionally expressive, and thus move up the recovery curve of the Uncanny Valley. Similarly, improved choices of avatar embodiment, fidelitous expressivity, and fulfilled proprioception would clearly help XR Telepresence technology move further up the recovery curve of the Uncanny Valley of Telepresence. That being said, there remains the question of just what combination of these features will form an acceptable gestalt in various collaboration contexts. We should not expect XR Telepresence to be one-size-fits-all. There will be a need for both specialised XR Telepresence systems for specific contexts and modular/flexible systems that can be adapted across changing contexts.

6.2 Uncertainties about the Other Partner's Capabilities

In our study, local and remote users had some uncertainties about what one another were capable of seeing, hearing, and expressing. Many of these are things that could be addressed through new additions to the interface. For one, both local and remote users were often unaware of how they were appearing in their partner's view. While we let remote participants wear the HoloLens and see what their avatar looked like on the local user's interface, remote participants were not always aware of how their gestures and head-gaze expressions appeared to their local partners. Similarly, local participants were not always aware of how well their remote partners could see into the space. Since participants did not know each other's capabilities, these had to be explicitly communicated. It would be beneficial to add elements that help users overcome and communicate these uncertainties. As a simple example, a feature that lets the remote users see their avatar could be helpful (e.g. as is possible in the Spatial system [11]). Another simple, yet potentially helpful solution would be to let the remote user peek into the local user's view through the HoloLens, or conversely, letting the local user peek into the remote user's view. To see the world from the other's point of view would provide a sense of reciprocity of perspectives that has eluded video-mediated communication for decades. This begins to address some of Hollan and Stornetta's [36] suggestions for features that treat communicative requirements, not media representations, as first-class citizens.

6.3 Limitations and Future Work

This study only looked at short-term usage of the experimental VROOM and standard Beam interfaces in an artificially-constructed game activity, and thus the findings only begin to scratch the surface.

The most obvious initial limitation was that VROOM was combined from existing technologies that were not intended to work together in this manner. Its seams, then, were very much on display. Future research improving the fidelity of each technology will clearly provide for better experiences, and future research that is able to develop holistic technologies will also clearly provide better comparisons to existing mature technologies. That being said, we also believe that flexibility should be a goal for future systems, so there is clearly also scope for research on how to enable modular systems that still feel holistic.

Similarly, the artificial nature of the task, as is often the case, does not speak to issues such as the need for either specialised or flexible XR Telepresence systems that can be fitted to their context. Future research could use more environmentally-valid tasks.

Given that VROOM was an experimental technology, and that even the standard Beam interface was new to some participants, some of our findings could have been due to novelty effects. People might adapt their behaviours or develop strategies over time, as they have been found to do in, for example, long term use of media spaces [19]. For example, remote users might become more used to the controls or the feeling of their embodiment from prolonged usage of VROOM, and thus might develop higher proprioception. Additionally, local users might end up doing less of the communicatively-limited mirroring of stillness and revert to more frequent gesturing as they become more used to VROOM.

The potential for adapting to VROOM highlights this study's limited time scale, which is exceptionally important not only to comfort with the technology, but the larger issue of hybrid collaboration in the workplace *over time*. Throughout a workday in a physical facility, we spend time in space together transitioning between multiple kinds of encounters held in different places. Some encounters are planned (e.g., meetings), some are serendipitous (e.g., hallway and water-cooler encounters), and some are framed by organisational location proximity (e.g., your nearest colleagues' desks). Each in the series of encounters has its own transitions into and out of the encounter itself, such as greeting a colleague as you both enter a meeting room, or making plans to follow-up as you exit the meeting room. All these transitions are themselves demonstrations of belonging. One can of course belong to a social group that is purely virtual, but one of the obvious values of offices and other places where people gather to conduct collaborative activities is that by coming to the space and moving around it over time, one is demonstrating membership in a set of nested and overlapping social groups. Membership is not just demonstrated by active engagement, as there is obviously an ambient demonstration and perception of presence that links personal autonomy to social affiliation.

While there will always be a need for lab studies with very novel technologies, we recommend that future work attend to issues of longer-term usage in real-world activities. Studies that look at habitual longitudinal usage on a day-to-day basis might reveal insights about how different technological approaches affect senses of 'belonging' that extend beyond what is physically perceived to what is both productively and emotionally engaging.

7 CONCLUSION

We are clearly still at the base of the steep far side of the Uncanny Valley of XR Telepresence. The VROOM technology probe was built using a combination of off-the-shelf technologies to provide a collaboration experience that explored what it might take to help remote and local users feel that a remote user belonged in a local activity space. Our study was exploratory, involving pairs of local and remote participants collaborating in a search game and playing a word-guessing game, comparing the standard MRP condition to VROOM. While VROOM did enable some features of embodiment, gesturing, mobility, spatial awareness, and non-verbal expressions, its limitations were clear. Challenges around proprioception, confusion regarding the mixing of a physical robot body with a virtual human avatar, uncertainties of one's partner's views and capabilities, the limited fidelity of the remote user's gestures and expressions, and the limited appearance of the avatar all contributed to remote users struggling to inhabit and exhibit a self that could act autonomously in the local activity space. Further, the more immersive an XR Telepresence system is, the more problems are amplified because of the higher expectations of both remote and local users. All that being said, if we put aside the pure goal of imitation, and concentrate more features that provide the comforts of belonging, even in unnatural ways, we may find some more traction in the valley. And, clearly, robot bodies and VR and MR headsets could be equipped with capabilities far beyond those of basic human face-to-face interaction. So not only should XR Telepresence seek new accountable ways of providing comfortable and contextual experiences, but they should aspire to provide capabilities that exceed those of humans. This, we argue, is what it will take to move Beyond Being There [36], to *Belonging There.*

ACKNOWLEDGMENTS

We thank our Microsoft colleagues who have helped us tremendously in developing the VROOM prototype: James Scott, Xu Cao, He Huang, Minnie Liu, Zhao Jun, Matthew Gan, and Leon Lu.

REFERENCES

[1] [n.d.]. Avatar Maker Pro - 3D avatar from a single selfie - Asset Store. https://assetstore.unity.com/packages/tools/modeling/avatar-maker-pro-3d-avatar-from-a-single-selfie-134800

[2] [n.d.]. Azure Kinect DK – Develop AI Models | Microsoft Azure. https://azure.microsoft.com/en-us/services/kinect-dk/ Library Catalog: azure.microsoft.com.

[3] [n.d.]. BEAM - From Here to Anywhere. https://suitabletech.com/

[4] [n.d.]. Double Robotics - Telepresence Robot for Telecommuters. https://www.doublerobotics.com/

[5] [n.d.]. Get Xbox Avatar Editor - Microsoft Store en-CA. https://www.microsoft.com/en-ca/p/xbox-avatar-editor/9nblggh4v0r3 Library Catalog: www.microsoft.com.

[6] [n.d.]. Hand Tracking. https://developer.oculus.com/documentation/unity/unity-handtracking/

[7] [n.d.]. Help Center - Desktop App. https://suitabletech.com/support/helpcenter/desktop-app-full-listing

[8] [n.d.]. HP Windows Mixed Reality Headset | HP® Official Site. https://www8.hp.com/us/en/campaigns/mixedrealityheadset/overview.html

[9] [n.d.]. Product | RICOH THETA V. https://theta360.com/en/about/theta/v.html

[10] [n.d.]. Spatial - Collaborate from anywhere in Augmented Reality. https://spatial.is/

[11] [n.d.]. Spatial - Collaborate from anywhere in Augmented Reality. https://spatial.io/

[12] 2020. microsoft/Microsoft-Rocketbox. https://github.com/microsoft/Microsoft-Rocketbox original-date: 2020-03-13T10:12:00Z.

[13] Ehsan Azimi, Long Qian, Nassir Navab, and Peter Kazanzides. 2018. Alignment of the Virtual Scene to the 3D Display Space of a Mixed Reality Head-Mounted Display. *arXiv preprint arXiv:1703.05834* (2018).

[14] Oriana Bandiera, Iwan Barankay, and Imran Rasul. 2008. Social capital in the workplace: Evidence on its formation and consequences. *Labour Economics* 15, 4 (Aug. 2008), 724–748. https://doi.org/10.1016/j.labeco.2007.07.006

[15] Frank Biocca. 1997. The Cyborg's Dilemma: Progressive Embodiment in Virtual Environments. *Journal of Computer-Mediated Communication* 3, 2 (Sept. 1997). https://doi.org/10.1111/j.1083-6101.1997.tb00070.x

[16] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.

[17] Andriana Boudouraki, Stuart Reeves, Joel E. Fischer, and Sean Rintel. 2020. "I can't get round": Recruiting Assistance in Mobile Robotic Telepresence. In *CSCW2020*. PACM HCI. https://www.microsoft.com/en-us/research/publication/i-cant-get-round-recruiting-assistance-in-mobile-robotic-telepresence/ Backup Publisher: ACM.

[18] Alphonse Chapanis, Robert B. Ochsman, Robert N. Parrish, and Gerald D. Weeks. 1972. Studies in Interactive Communication: I. The Effects of Four Communication Modes on the Behavior of Teams During Cooperative Problem-Solving. *Human Factors* 14, 6 (Dec. 1972), 487–509. https://doi.org/10.1177/001872087201400601

[19] Paul Dourish, Annette Adler, Victoria Bellotti, and Austin Henderson. 1996. Your place or mine? Learning from long-term use of audio-video communication. *Computer Supported Cooperative Work (CSCW)* 5, 1 (1996), 33–62. Publisher: Springer.

[20] Paul W. Eastwick and Wendi L. Gardner. 2009. Is it a game? Evidence for social influence in the virtual world. *Social Influence* 4, 1 (Jan. 2009), 18–32. https://doi.org/10.1080/15534510802254087

[21] Omid Fakourfar, Kevin Ta, Richard Tang, Scott Bateman, and Anthony Tang. 2016. Stabilized Annotations for Mobile Remote Assistance. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1548–1560. https://doi.org/10.1145/2858036.2858171 event-place: San Jose, California, USA.

[22] Kathleen E. Finn, Abigail J. Sellen, and Sylvia B. Wilbur. 1997. *Video-Mediated Communication.* L. Erlbaum Associates Inc., USA.

[23] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/358916.358947 event-place: Philadelphia, Pennsylvania, USA.

[24] Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. 2004. Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Human–Computer Interaction* 19, 3 (Sept. 2004), 273–309. https://doi.org/10.1207/s15327051hci1903_3 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/s15327051hci1903_3.

[25] Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias Höllerer. 2012. Integrating the Physical Environment into Mobile Remote Collaboration. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '12)*. ACM, New York, NY, USA, 241–250. https://doi.org/10.1145/2371574.2371610 event-place: San Francisco, California, USA.

[26] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-stabilized Annotations and Virtual Scene Navigation for Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 449–459. https://doi.org/10.1145/2642918.2647372 event-place: Honolulu, Hawaii, USA.

[27] William Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is not enough: multiple views in a media space. In *Proceedings of the INTERCHI '93 conference on Human factors in computing systems (INTERCHI '93)*. IOS Press, Amsterdam, The Netherlands, 335–341.

[28] William W. Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is Not Enough: Multiple Views in a Media Space. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 335–341. https://doi.org/10.1145/169059.169268 event-place: Amsterdam, The Netherlands.

[29] James Paul Gee. 2008. Video Games and Embodiment. *Games and Culture* 3, 3-4 (July 2008), 253–263. https://doi.org/10.1177/1555412008317309

[30] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, Angus Antley, Anthony Steed, Bernhard Spanlang, Antonella Maselli, Domna Banakou, Nuria Pelechano, Sergio Orts-Escolano, Veronica Orvalho, Laura Trutoiu, Markus Wojcik, Maria V. Sanchez-Vives, Jeremy Bailenson, Mel Slater, and Jaron Lanier. 2020. The Rocketbox Library and the Utility of Freely Available Rigged Avatars. *Frontiers in Virtual Reality* 1 (2020). https://doi.org/10.3389/frvir.2020.561558 Publisher: Frontiers.

[31] Mar González-Franco, Daniel Pérez-Marcos, Bernhard Spanlang, and Mel Slater. 2010. The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment. In *2010 IEEE Virtual Reality Conference (VR)*. 111–114. https://doi.org/10.1109/VR.2010.5444805 ISSN: 2375-5334.

[32] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (Sept. 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[33] Carl Gutwin and Reagan Penner. 2002. Improving Interpretation of Remote Gestures with Telepointer Traces. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW '02)*. ACM, New York, NY, USA, 49–57. https://doi.org/10.1145/587078.587086 event-place: New Orleans, Louisiana, USA.

[34] Christian Heath and Paul Luff. 1992. Media space and communicative asymmetries: preliminary observations of video-mediated interaction. *Human-Computer Interaction* 7, 3 (Dec. 1992), 315–346. https://doi.org/10.1207/s15327051hci0703_3

[35] Yasamin Heshmat, Brennan Jones, Xiaoxuan Xiong, Carman Neustaedter, Anthony Tang, Bernhard E. Riecke, and Lillian Yang. 2018. Geocaching with a Beam: Shared Outdoor Activities Through a Telepresence Robot with 360 Degree Viewing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 359:1–359:13. https://doi.org/10.1145/3173574.3173933 event-place: Montreal QC, Canada.

[36] Jim Hollan and Scott Stornetta. 1992. Beyond Being There. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. ACM, New York, NY, USA, 119–125. https://doi.org/10.1145/142750.142769 event-place: Monterey, California, USA.

[37] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[38] Wijnand A. IJsselsteijn, Jonathan Freeman, and Huib De Ridder. 2001. *Presence: Where are we?* Mary Ann Liebert, Inc.

[39] Brennan Jones, Kody Dillman, Richard Tang, Anthony Tang, Ehud Sharlin, Lora Oehlberg, Carman Neustaedter, and Scott Bateman. 2016. Elevating Communication, Collaboration, and Shared Experiences in Mobile Video Through Drones. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. ACM, New York, NY, USA, 1123–1135. https://doi.org/10.1145/2901790.2901847

[40] Brennan Jones, Anna Witcraft, Scott Bateman, Carman Neustaedter, and Anthony Tang. 2015. Mechanics of Camera Work in Mobile Video Collaboration. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 957–966. https://doi.org/10.1145/2702123.2702345

[41] Brennan Jones, Yaying Zhang, Priscilla N. Y. Wong, and Sean Rintel. 2020. VROOM: Virtual Robot Overlay for Online Meetings. In *Extended Abstracts of the 2020 ACM Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3334480.3382820

[42] Brigitte Jordan and Austin Henderson. 1995. Interaction Analysis: Foundations and Practice. *Journal of the Learning Sciences* 4, 1 (Jan. 1995), 39–103. https://doi.org/10.1207/s15327809jls0401_2 Publisher: Routledge _eprint: https://doi.org/10.1207/s15327809jls0401_2.

[43] Shunichi Kasahara and Jun Rekimoto. 2015. JackIn Head: Immersive Visual Telepresence System with Omnidirectional Wearable Camera for Remote Collaboration. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology (VRST '15)*. ACM, New York, NY, USA, 217–225. https://doi.org/10.1145/2821592.2821608

[44] Anuroop Katiyar, Karan Kalra, and Chetan Garg. 2015. Marker based augmented reality. *Advances in Computer Science and Information Technology (ACSIT)* 2, 5 (2015), 441–445.

[45] Mohammad Keshavarzi, Woojin Ko, Allen Y. Yang, and Luisa Caldas. 2020. Optimization and Manipulation of Contextual Mutual Spaces for Multi-User Virtual and Augmented Reality Interaction. *arXiv preprint arXiv:1910.05998* (2020).

[46] Konstantina Kilteni, Ilias Bergstrom, and Mel Slater. 2013. Drumming in Immersive Virtual Reality: The Body Shapes the Way We Play. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (April 2013), 597–605. https://doi.org/10.1109/TVCG.2013.29 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[47] Konstantina Kilteni, Jean-Marie Normand, Maria V. Sanchez-Vives, and Mel Slater. 2012. Extending Body Space in Immersive Virtual Reality: A Very Long Arm Illusion. *PLOS ONE* 7, 7 (July 2012), e40867. https://doi.org/10.1371/journal.pone.0040867 Publisher: Public Library of Science.

[48] Seungwon Kim, Sasa Junuzovic, and Kori Inkpen. 2014. The Nomad and the Couch Potato: Enriching Mobile Shared Experiences with Contextual Information. In *Proceedings of the 18th International Conference on Supporting Group Work*. ACM, 167–177.

[49] David Kirk, Andy Crabtree, and Tom Rodden. 2005. Ways of the Hands. In *ECSCW 2005*, Hans Gellersen, Kjeld Schmidt, Michel Beaudouin-Lafon, and Wendy Mackay (Eds.). Springer Netherlands, 1–21. https://doi.org/10.1007/1-4020-4023-7_1

[50] David Kirk and Danae Stanton Fraser. 2006. Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1191–1200. https://doi.org/10.1145/1124772.1124951 event-place: Montréal, Québec, Canada.

[51] Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: a process model. In *Language and Gesture*, DavidEditor McNeill (Ed.). Cambridge University Press, 261–283. https://doi.org/10.1017/CBO9780511620850.017

[52] Robert E. Kraut, Susan R. Fussell, Susan E. Brennan, and Jane Siegel. 2002. Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. *Distributed work* (2002), 137–162.

[53] Kwan Min Lee. 2004. Presence, Explicated. *Communication Theory* 14, 1 (Feb. 2004), 27–50. https://doi.org/10.1111/j.1468-2885.2004.tb00302.x

[54] Min Kyung Lee and Leila Takayama. 2011. "Now, i have a body": uses and social norms for mobile remote presence in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, Vancouver, BC, Canada, 33–42. https://doi.org/10.1145/1978942.1978950

[55] Christian Licoppe, Paul K. Luff, Christian Heath, Hideaki Kuzuoka, Naomi Yamashita, and Sylvaine Tuncer. 2017. Showing Objects: Holding and Manipulating Artefacts in Video-mediated Collaborative Settings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, Denver,

Colorado, USA, 5295–5306. https://doi.org/10.1145/3025453.3025848

[56] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication* 3, 2 (Sept. 1997). https://doi.org/10.1111/j.1083-6101.1997.tb00072.x

[57] Paul Luff, Christian Heath, Hideaki Kuzuoka, Jon Hindmarsh, Keiichi Yamazaki, and Shinya Oyama. 2003. Fractured ecologies: creating environments for collaboration. *Human-Computer Interaction* 18, 1 (June 2003), 51–84. https://doi.org/10.1207/S15327051HCI1812_3

[58] Marvin Minsky. 1980. Telepresence. (1980).

[59] Masahiro Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.

[60] Joan Mulholland. 1996. A series of story turns: Intertextuality and collegiality. *Text & Talk* 16, 4 (Dec. 1996), 535–556. https://doi.org/10.1515/text.1.1996.16.4.535 Publisher: De Gruyter Mouton Section: Text & Talk.

[61] Thomas Neumayr, Hans-Christian Jetter, Mirjam Augstein, Judith Friedl, and Thomas Luger. 2018. Domino: A Descriptive Framework for Hybrid Collaboration and Coupling Styles in Partially Distributed Teams. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 128:1–128:24. https://doi.org/10.1145/3274397

[62] Carman Neustaedter, Samarth Singhal, Rui Pan, Yasamin Heshmat, Azadeh Forghani, and John Tang. 2018. From Being There to Watching: Shared and Dedicated Telepresence Robot Usage at Academic Conferences. *ACM Transactions on Computer-Human Interaction* 25, 6 (Dec. 2018), 33:1–33:39. https://doi.org/10.1145/3243213

[63] Carman Neustaedter, Gina Venolia, Jason Procyk, and Daniel Hawkins. 2016. To Beam or Not to Beam: A Study of Remote Telepresence Attendance at an Academic Conference. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16).* ACM, New York, NY, USA, 418–431. https://doi.org/10.1145/2818048.2819922

[64] Kenton O'Hara, Alison Black, and Matthew Lipson. 2006. Everyday Practices with Mobile Video Telephony. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06).* ACM, New York, NY, USA, 871–880. https://doi.org/10.1145/1124772.1124900 event-place: Montréal, Québec, Canada.

[65] Tuula Oksanen, Ichiro Kawachi, Anne Kouvonen, Soshi Takao, Etsuji Suzuki, Marianna Virtanen, Jaana Pentti, Mika Kivimäki, and Jussi Vahtera. 2013. Workplace Determinants of Social Capital: Cross-Sectional and Longitudinal Evidence from a Finnish Cohort Study. *PLOS ONE* 8, 6 (June 2013), e65846. https://doi.org/10.1371/journal.pone.0065846 Publisher: Public Library of Science.

[66] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics* 35, 6 (Nov. 2016), 221:1–221:14. https://doi.org/10.1145/2980179.2980252

[67] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16).* Association for Computing Machinery, Tokyo, Japan, 741–754. https://doi.org/10.1145/2984511.2984517

[68] Jiazhi Ou, Susan R. Fussell, Xilin Chen, Leslie D. Setlock, and Jie Yang. 2003. Gestural communication over video stream: supporting multimodal interaction for remote collaborative physical tasks. In *Proceedings of the 5th international conference on Multimodal interfaces (ICMI '03).* Association for Computing Machinery, New York, NY, USA, 242–249. https://doi.org/10.1145/958432.958477

[69] Eric Paulos and John Canny. 1998. PRoP: Personal Roving Presence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98).* ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 296–303. https://doi.org/10.1145/274644.274686

[70] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16).* ACM, New York, NY, USA, 1716–1725. https://doi.org/10.1145/2818048.2819965 event-place: San Francisco, California, USA.

[71] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, 46:1–46:13. https://doi.org/10.1145/3173574.3173620 event-place: Montreal QC, Canada.

[72] Thammathip Piumsomboon, Gun A. Lee, Andrew Irlitti, Barrett Ens, Bruce H. Thomas, and Mark Billinghurst. 2019. On the Shoulder of the Giant: A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, 228:1–228:17. https://doi.org/10.1145/3290605.3300458 event-place: Glasgow, Scotland Uk.

[73] Long Qian. 2019. qian256/HoloLensARToolKit. https://github.com/qian256/HoloLensARToolKit original-date: 2017-01-19T20:04:54Z.

[74] Irene Rae and Carman Neustaedter. 2017. Robotic Telepresence at Scale. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 313–324. https://doi.org/10.1145/3025453.3025855

[75] Irene Rae, Gina Venolia, John C. Tang, and David Molnar. 2015. A Framework for Understanding and Designing Telepresence. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1552–1566. https://doi.org/10.1145/2675133.2675141 event-place: Vancouver, BC, Canada.

[76] Frances H Rauscher, Robert M Krauss, and Yihsiu Chen. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological science* 7, 4 (1996), 226–231. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[77] Taehyun Rhee, Stephen Thompson, Daniel Medeiros, Rafael Dos Anjos, and Andrew Chalmers. 2020. Augmented Virtual Teleportation for High-Fidelity Telecollaboration. *IEEE Transactions on Visualization and Computer Graphics* (2020). Publisher: IEEE.

[78] Banu Saatçi, Kaya Akyüz, Sean Rintel, and Clemens Nylandsted Klokmose. 2020. (Re)Configuring Hybrid Meetings: Moving from User-Centered Design to Meeting-Centered Design. *Computer Supported Cooperative Work (CSCW): The Journal of Collaborative Computing and Work Practices* (2020).

[79] Banu Saatçi, Roman Rädle, Sean Rintel, Kenton O'Hara, and Clemens Nylandsted Klokmose. 2019. Hybrid Meetings in the Modern Workplace: Stories of Success and Failure. In *Collaboration Technologies and Social Computing (Lecture Notes in Computer Science)*, Hideyuki Nakanishi, Hironori Egi, Irene-Angelica Chounta, Hideyuki Takada, Satoshi Ichimura, and Ulrich Hoppe (Eds.). Springer International Publishing, Cham, 45–61. https://doi.org/10.1007/978-3-030-28011-6_4

[80] Lilyan Salazar. [n.d.]. Beam supports Microsoft Research's efforts to save costs and time. https://suitabletech.com/casestudy-microsoft Library Catalog: suitabletech.com.

[81] scooley. [n.d.]. Getting around HoloLens 2. https://docs.microsoft.com/en-us/hololens/hololens2-basic-usage Library Catalog: docs.microsoft.com.

[82] Abigail Sellen, Bill Buxton, and John Arnott. 1992. Using spatial cues to improve videoconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. Association for Computing Machinery, Monterey, California, USA, 651–652. https://doi.org/10.1145/142750.143070

[83] Hanieh Shakeri and Carman Neustaedter. 2019. Teledrone: Shared Outdoor Exploration Using Telepresence Drones. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. Association for Computing Machinery, Austin, TX, USA, 367–371. https://doi.org/10.1145/3311957.3359475

[84] Thomas B. Sheridan. 1992. Musings on telepresence and virtual presence. *Presence: Teleoperators & Virtual Environments* 1, 1 (1992), 120–126.

[85] Leila Takayama and Helen Harris. 2013. Presentation of (telepresent) self: On the double-edged effects of mirrors. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 381–388. https://doi.org/10.1109/HRI.2013.6483613 ISSN: 2167-2148.

[86] Anthony Tang, Omid Fakourfar, Carman Neustaedter, and Scott Bateman. 2017. Collaboration with 360° Videochat: Challenges and Opportunities. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, Edinburgh, United Kingdom, 1327–1339. https://doi.org/10.1145/3064663.3064707

[87] Katherine M. Tsui, Munjal Desai, Holly A. Yanco, and Chris Uhlik. 2011. Exploring use cases for telepresence robots. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 11–18. https://doi.org/10.1145/1957656.1957664 ISSN: 2167-2148.

[88] Roel Vertegaal. 1997. Conversational awareness in multiparty VMC. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems (CHI EA '97)*. Association for Computing Machinery, Atlanta, Georgia, 6–7. https://doi.org/10.1145/1120212.1120217

[89] Bin Xu, Jason Ellis, and Thomas Erickson. 2017. Attention from Afar: Simulating the Gazes of Remote Participants in Hybrid Meetings. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, Edinburgh, United Kingdom, 101–113. https://doi.org/10.1145/3064663.3064720

[90] Lillian Yang, Brennan Jones, Carman Neustaedter, and Samarth Singhal. 2018. Shopping Over Distance Through a Telepresence Robot. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 191:1–191:18. https://doi.org/10.1145/3274460

[91] Lillian Yang and Carman Neustaedter. 2018. Our House: Living Long Distance with a Telepresence Robot. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–18. https://doi.org/10.1145/3274459

[92] Lillian Yang and Carman Neustaedter. 2020. An Autobiographical Design Study of a Long Distance Relationship: When Telepresence Robots Meet Smart Home Tools. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM Press, New York, NY, USA. http://clab.iat.sfu.ca/pubs/Yang-RobotsSmartDevices-DIS2020.pdf

[93] Lillian Yang, Carman Neustaedter, and Thecla Schiphorst. 2017. Communicating Through A Telepresence Robot: A Study of Long Distance Relationships. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors*

in Computing Systems (CHI EA ’17). ACM, New York, NY, USA, 3027–3033. https://doi.org/10.1145/3027063.3053240

A INTERVIEW QUESTIONS

(1) How do you feel about the experience working with your partner?
(2) What aspects of the experience worked well for you?
(3) **(If they did not mention VR/AR experience,)**
  (a) **[For RUs]** Can you describe your experience using the telepresence robot or VR view?
  (b) **[For LUs]** Can you describe your experience seeing and interacting with your partner's telepresence robot or the avatar?
(4) What aspects of the experience did not work well for you?
(5) **(If they did not mention VR/AR experience,)**
  (a) **[For RUs]** Can you describe your experience using the telepresence robot or VR view?
  (b) **[For LUs]** Can you describe your experience seeing and interacting with your partner's telepresence robot or the avatar?
(6) What does this feel like vs. regular video conferencing?
(7) Where there any specific strategies that you employed to communicate/collaborate with your partner? / Did you do certain things to help with the communication with your partner?
  (a) **[For RUs]** Did you employ different strategies when you were wearing a VR headset versus when you were just looking at the screen?
  (b) **[For LUs]** Did these strategies differ when you were able to see your partner's avatar over the robot vs. when you were not able to?
(8) Were there any moments when you felt like you were able to understand each other well? [if yes,] Can you describe the situation and how you felt.
  (a) What do you think led to your partner successfully understanding you?
(9) Were there any moments when you felt like you were not able to understand each other well? [if yes,] Can you describe the situation and how you felt.
  (a) How did you try to avoid and/or fix misunderstandings?
(10) How much do you identify with your avatar (both appearance and body)?