Int. J. Human–Computer Studies 205 (2025) 103632

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Avatars in mixed-reality meetings: A longitudinal field study of realistic versus cartoon facial likeness effects on communication, task satisfaction, presence, and emotional perception

Georgiana Cristina Dobre [a,b,c,*], Marta Wilczkowiak [c], Marco Gillies [b], Xueni Pan [b], Sean Rintel [c]

[a] Human-Centered Artificial Intelligence, University of Augsburg, Augsburg, Germany
[b] SEEVR Lab, Department of Computing, Goldsmiths, University of London, London, United Kingdom
[c] Microsoft Research, Cambridge, United Kingdom

ARTICLE INFO

Keywords:
Avatar rendering style
Facial realism
Mixed reality
Meetings
Work
Longitudinal field study
Ecological validity
Communication
Presence
Tasks
Emotion

ABSTRACT

We conducted a within-subjects study to examine how realistic faces and cartoon faces on avatars affect communication, task satisfaction, sense of presence, and mood perception in mixed reality meetings. Over the course of two weeks, six groups of co-workers (14 people) held recurring meetings using Microsoft HoloLens2 devices, each person embodying a personal full-body avatar with either a realistic face or cartoon face. Half of the groups started with the realistic face avatar and switched to the cartoon face version halfway through (RC condition), and the other half with the cartoon-face avatar first (CR condition). Results showed that participants in the RC condition may have had higher expectations and more errors in perceiving their colleagues' moods. Participants in the CR condition reported that the avatars' appearance mattered less over time and experienced increased comfort and improved identification of their colleagues. Participants rated words, tone of voice, and movement as the most useful cues for perceiving colleagues' moods, regardless of avatar rendering style. In the RC condition, participants rated gaze as more useful than facial expressions, while in the CR condition, both gaze and facial expressions were rated as the least useful. Results also suggested that participants had more errors when perceiving negative moods in their colleagues, with this trend appearing for most moods, but depending on conditions. Implications of these findings for mixed and virtual reality meetings are discussed. This work contributes to the field of remote collaboration by providing insights from longitudinal data on the impact of avatar appearance on various aspects of work meetings in virtual environments.

1. Introduction

As remote meetings have facilitated a significant increase in global collaboration, there has been a growing demand for $3D$ immersive systems that address the limitations of traditional $2D$ formats. The aim of these $3D$ systems is to connect remote users as if they were in the same location. This allows people to work more effectively on shared tasks because the value of mixed and virtual reality (MR/VR) meetings is the preservation of spatial relationships (Otto et al., 2006) and social behaviours such as proximity or gaze (Bailenson et al., 2001).

When people use this technology for remote collaborations, they embody an avatar. These avatars represent people's identities, positions, interests, and activities (Benford et al., 1995). Avatars can have different representations, ranging from floating spheres with hands to full or partial humanoid bodies with different appearance rendering styles (e.g., cartoon, realistic). Thanks to advances in technology, avatars can be highly customised to resemble a person and follow a particular style. There are positive and negative aspects to different avatar rendering styles. For example, the use of realistic avatars may make people feel uncomfortable and lower their feelings of affinity (Shin et al., 2019). This is often due to the discrepancy between high expectations of nonverbal behaviour (such as body movement, facial expressions) and the avatar's actual behaviour. Cartoon rendering style, whether generic or customised, may lead to anxious feelings about the appropriateness of non-realistic representations in a professional context (Bailenson and Beall, 2006). Most of the research on avatars focuses on presence, workload, or trust (Waltemate et al., 2018; Lugrin et al., 2015; Latoschik et al., 2017; Yoon et al., 2019; Khojasteh and Won, 2021; Heidicker et al., 2017), with mixed results (see Section 2).

Furthermore, during studies, participants often look at only short animations or still images of avatars (MacDorman and Chattopadhyay,

* Corresponding author at: Human-Centered Artificial Intelligence, University of Augsburg, Augsburg, Germany.
E-mail address: cristina.dobre@uni-a.de (G.C. Dobre).

https://doi.org/10.1016/j.ijhcs.2025.103632
Received 2 August 2024; Received in revised form 6 September 2025; Accepted 9 September 2025
Available online 30 September 2025

2016; Shin et al., 2019) and/or have one-off interactions with others (Lugrin et al., 2015; Waltemate et al., 2018; Jo et al., 2017; Yoon et al., 2019; Zibrek et al., 2018; Heidicker et al., 2017), making the findings prone to novelty effects (Koch et al., 2018; Parmar, 2017). However, real-life collaborative work in immersive environments involves users who know each other and interact regularly, trying to get real work done. The *communicative functionality* of avatars is essential in these cases. Since the spatial audio common to most immersive environments provides a highly naturalistic vocal representation, it is the *nonverbal* communicative functionality that is primarily at issue, such as the ability to identify each other, recognise facial expressions and gestures (Burgoon et al., 2016), negotiate proxemics (Hall et al., 1968), and, when presented virtually, trust that these are authentic representations of their colleagues (Oh et al., 2018).

Finally, there is a small body of longitudinal research investigating how the impact of avatar appearance on participants' behaviour, attitudes, and interactions changes over time. Bailenson and Yee (2006) and Han et al. (2023) found that in avatar-mediated structured group interactions, task performance, subjective ratings, nonverbal behaviour, entitativity, presence, enjoyment, realism, and synchrony changed over time. Brown et al. (2024) and Latifi et al. (2024) demonstrated user–avatar bond variation over time in the context of gaming. To our knowledge, there is no study investigating the effects of avatar realism in ecologically valid workplace conditions.

In summary, most research on avatar appearance in meeting settings comes from one-off lab studies in virtual reality environments. We know little about how these findings apply to MR, less about effects in real-world contexts, and very little about the longitudinal effects on avatar acceptance. To our knowledge, there is a gap in VR/MR literature regarding this combination of aspects.

This research addresses that gap by investigating how people feel about using avatars with different appearance styles in immersive meetings over multiple sessions. For two to three weeks, six groups of co-workers (14 people) from a global technology company conducted a series of virtual meetings using Microsoft HoloLens2 (HL2) devices. Each participant used a personalised avatar with either a realistic or cartoon face. Half the groups began with realistic avatars and the other half with cartoon avatars; all groups switched halfway through the study period. Our main focus was to determine whether the acceptance ratings for both realistic and cartoon avatars would change over time as novelty waned. Specifically, we were interested in the functional communicative value, task satisfaction, presence, and the self-reported and perceived moods of individuals during immersive virtual meetings. In the following sections, we report on relevant prior work (Section 2), introduce the research questions (Section 3) and detail our methodology (Section 4). We then present our data analysis (Section 4.6) and results (Section 5), before embarking on a discussion of these findings (Section 6). Finally, we offer a conclusion that outlines potential avenues for future work (Section 8).

2. Related work

The appearance, body representation, and resemblance of an avatar all play a crucial role in determining the level of trust, efficiency, and presence experienced during virtual interactions. Prior research has shown that the use of avatars can enhance these aspects of social interaction in Immersive Virtual Environments (IVEs), comparable to face-to-face interactions (Yoon et al., 2019; Pan and Steed, 2017). On the other hand, forgoing an avatar or being represented only by hands or controllers can lead to a deterioration of communication and feelings of loneliness among participants (Smith and Neff, 2018). In this section, we delve into the related work on these factors and the importance of temporality in the field of social interactions in IVEs.

Avatar rendering style

Avatar rendering style can significantly impact the way users experience and perceive VR environments. The appearance of an avatar can greatly influence users' sense of embodiment, social presence, and trust (Pan and Steed, 2017; Smith and Neff, 2018; Collingwoode-Williams et al., 2021).

The most common way to represent avatars are in a realising or in a cartoon rendering style (see Weidner et al. (2023)'s systematic review on avatars' rendering in virtual environments). Realistic representations include avatars created via 3D modelling, 3D scanning, video-avatars (streaming the 2D video of a user to IVEs), or point-cloud avatars. Realistic representations are both more difficult to create and may evoke the uncanny valley effect (Lugrin et al., 2015). This gives room to cartoon style avatars as they are more stylised and simplified.

Literature focused on different avatar rendering styles found no significant differences between them (Yoon et al., 2019; Garcia et al., 2021; Fraser et al., 2024), or contradictory results (Pakanen et al., 2022; Sakurai et al., 2021).

On one hand, research shows participants preferring realistic avatars (Yuan et al., 2019; Pakanen et al., 2022; Arboleda et al., 2024), reporting higher quality of experience while using them (De Simone et al., 2019), or higher social presence and higher attractiveness ratings (Amadou et al., 2023). Salagean et al. (2023) found that highly photorealistic and personalised avatars increased their embodiment, self-identification and had positive avatar perception effects. Latoschik et al. (2017) found that participants reported higher body ownership when using realistic avatars compared to wooden-block-person ones.

On the other hand, realistic appearance have been reported as lacking in human spark ("their eyes seemed empty") and lacking communicative flexibility ("expressions were hard to read") (Sakurai et al., 2021), illustrating the issue of the uncanny valley effect. Lugrin et al. (2015) conducted an experiment in which they compared the effects of using a robot avatar, a block-person avatar, and a realistic avatar in a find-and-touch game set in a virtual forest environment. They found that the use of realistic avatars led to a lower illusion of virtual body ownership. McDonnell et al. (2012) compared different character render styles from short video clips, with the realistic rendering being preferred over the cartoon one. However, when large motion artefacts were present, participants considered the cartoon rendering more appealing and more pleasant than the realistic one. At the same time, Sonia et al. (2023) argue that a higher level of visual detail for facial expressivity is essential for avoiding uncanny valley in VCs.

In addition to appearance, personality can also play a role in users' affinity towards a virtual character (VC). Zibrek et al. (2018) found that the user's affinity towards the VC was based on the VC's appearance and personality, and that realism in VC's appearance can be a positive choice in VR. Furthermore, Suma et al. (2023) discovered a link between facial expression intensity and the emotion recognition.

Avatar body

Several studies have examined the effects of various body structures on VR and MR experiences, including full-body, upper-body, head and hands, and controller-only avatars (Yoon et al., 2019; Pan and Steed, 2017; Smith and Neff, 2018; Herrera et al., 2018; Aseeri and Interrante, 2021; Collingwoode-Williams et al., 2021; Pakanen et al., 2022). In general, participants preferred full-body avatars, which were associated with higher levels of social presence and co-presence (Yoon et al., 2019; Smith and Neff, 2018; Aseeri and Interrante, 2021), increased trust and faster task completion (Pan and Steed, 2017), and overall higher preference (Aseeri and Interrante, 2021). Pan and Steed (2017), Smith and Neff (2018) compared head-and-hands avatars to full-body cartoon or robot style avatars. In both cases, the full-body avatar was preferred against the simplistic head-and-hands representation, showing higher levels of social presence and trust. Similarly, in two surveys with 16 and 87 participants respectively, Pakanen et al. (2022) asked participants to rank their first, second, and third preference for avatar appearance

in VR and MR. Participants chose from 36 pictures of avatars with different representation styles. The preferred avatar for both VR and MR was the realistic full-body avatar, with the full-body 'hologram' avatar (which had a translucent alpha effect) being the second most popular choice for MR.

However, this does not always hold. Herrera et al. (2018) conducted a study in which only the movements of the hands and head were mapped from participants' movements, with all other body parts remaining static. Participants using head-and-hands avatars demonstrated higher social presence, self-presentation, and interpersonal attraction compared to those using full-body avatars in the same cartoon rendering style.

Avatar–user resemblance

In many cases, researchers have recruited groups of participants who are unfamiliar with one another and observed their experiences while using pre-defined avatars (Yoon et al., 2019; Pan and Steed, 2017; Smith and Neff, 2018; Aseeri and Interrante, 2021; Freiwald et al., 2021; Sakurai et al., 2021; Pakanen et al., 2022). However, this may not accurately reflect how avatars are (or will be) used in real life, as people often interact with acquaintances while using avatars in virtual environments.

In Kim et al. (2023)'s work, participants played a shooter VR game while embodying avatars that resemble themselves or other people of the same gender. Their results show that avatar self-similarity increases users' embodiment and social presence, though it has low effects on overall presence and slightly lowers immersion. Additionally, they found voice to contribute the most to the avatars' self-similarity, surpassing other representational factors.

Moustafa and Steed (2018) conducted an experiment in which they provided 9 groups of friends or family with VR headsets and asked them to meet in VR regularly for a month. Participants were able to customise their avatars using the options available in the GearVR application. The researchers found that participants were influenced by the group dynamics to adjust their avatar appearance to fit a version that resembled them. De Simone et al. (2019) had dyads of acquainted participants embody both customised cartoon avatars and personalised realistic avatars (via video stream). They asked the participants to watch a video together in VR and rate the quality of their experience in comparison to watching a video together in person. The personalised realistic avatars received ratings that were similar to those given for the in-person condition, whereas the experience quality using cartoon avatars was rated as the lowest.

Tasks and environment setting

Many studies have looked at the impact of avatar appearance on various tasks, such as playing games (Yoon et al., 2019; Moustafa and Steed, 2018; Khojasteh and Won, 2021; Langa et al., 2022; Pan and Steed, 2017; Smith and Neff, 2018; Herrera et al., 2018; Aseeri and Interrante, 2021; Pakanen et al., 2022) or tasks requiring more movement (Lugrin et al., 2015; Freiwald et al., 2021; Sakurai et al., 2021). Other tasks that have been examined include listening tasks (Zibrek et al., 2018; Yuan et al., 2019; Garcia et al., 2021), waving in a mirror (Latoschik et al., 2017), and watching videos (De Simone et al., 2019). However, fewer studies have focused on more formal tasks that typically take place in professional settings, such as brainstorming (Sun and Won, 2021), work meetings, conference networking (Nordin Forsberg and Kirchner, 2021), or classroom work and discussion (Han et al., 2022).

Nordin Forsberg and Kirchner (2021) explored the use of avatars in virtual business contexts using semi-structured interviews with two groups of participants: conference attendees using customised realistic avatars and coworkers using personalised realistic avatars in a VR business meeting. The researchers found that the participants in the meeting did not feel restricted by the appearance of their avatars. In the conference scenario, participants reported that the avatars helped them "break the ice" and initiate conversation, but also mentioned difficulties in recognising different people.

Sun and Won (2021) conducted a study in which dyads of participants completed a brainstorming task in VR while using either a personalised realistic avatar or a cube avatar. The participants were strangers to each other. After the task, they were asked about their own emotional state and the perceived emotional state of their partner. The researchers did not find any differences in emotional state recognition between the two different avatars.

Temporality in IVEs communication

Previous research on longitudinal studies in IVEs has shown that they are more ecologically valid and can provide insights into user behaviour changes, but they tend to take more time and resources to conduct. For example, in a study by Bailenson and Yee (2006), participants experienced less simulation sickness, had a stronger connection with their team over time, but they did not report significant changes in the level of presence and co-presence. Moustafa and Steed (2018) report that friends and family members who met in GearVR 1–2 times per week for a month updated their avatars to resemble themselves more accurately over time, at the request of others who found the interactions with the initial avatar uncomfortable and unnatural. The VR environment did not allow for nonverbal behaviours or facial expressions, as the current technology might not be accurate enough for complex emotions and facial movements (Hartbrich et al., 2023). Thus, initially participants had difficulty interpreting social cues. However, over time, they learned to rely on other cues such as voice tone.

Khojasteh and Won (2021) conducted a longitudinal study in Facebook Spaces where participants in dyads met for 5 sessions and played games in VR. Over time, participants became more comfortable using the controllers and the app, which allowed them to better connect with their partners. Again, since the system did not implement facial expressions, the participants also learned to use voice tone and word choice to perceive their partner's emotional state. Some participants reported improvements over time in completing tasks, but there was no significant difference in workload over time.

Han et al. (2022) conducted a longitudinal study which compared customised avatars to generic avatars. Eighty-one students participated in 8 weekly discussion sessions in the Engage VR platform, alternating between using platform-customised avatars and uniform upper-body avatars (bald avatars in school uniform clothing). Using growth models, the researchers found improvements over time in presence, enjoyment, entitativity, and realism. Groups that knew each other prior to the study showed higher social presence and enjoyment. Participants using the generic avatars reported lower self-presence but higher levels of enjoyment.

In sum, prior research results on the effects of realism versus cartoon styling of avatars are decidedly mixed, and they depend a great deal on the context and timing of participants' engagement. It seems that for body appearance there is a greater likelihood of preference for full-body traditionally-proportioned "realistic" avatar bodies compared to heads-and-hands, robot, block-style, or non-traditionally-proportioned cartoon style bodies (Aseeri and Interrante, 2021; Yoon et al., 2019; Pan and Steed, 2017; Smith and Neff, 2018; Herrera et al., 2018; Pakanen et al., 2022). Results on realistic versus cartoon styling in facial appearance are less clear cut, as are the interactions with gestural capabilities of traditionally-proportioned "realistic" avatar bodies, especially over time and engaged in real-world tasks.

To the best of our knowledge, there have been no studies comparing personalised realistic versus cartoon face styles on the same full-body avatars in IVEs, over time, in the field, and especially in business contexts. Similarly, very few studies focus holistically on the communicative encounter—the functional communicative value, task satisfaction, sense of presence, and the self-reported and perceived emotional states of the individuals (Nordin Forsberg and Kirchner, 2021; Garcia et al., 2021; Sun and Won, 2021). Given the emergent popularity of meetings in IVEs, and the likely variety of choices that IVEs will provide users, it is crucial to compare and contrast the experiences afforded by realistic and cartoon styling.

Table 1
Details on the groups size, participants' demographic, avatar order, sessions and questionnaires: the 12-item Likert-scale questionnaire (*12 it.* in table), the self reported moods (*Self*), the colleagues' perceived moods (*Perc.*), and the most useful cues) (*Cues*). There are 18(*) (instead of 20) questionnaires sets filled in for group *G*2 because, due to a technical error, there is a missing set of questionnaires from the last session using the cartoon avatars. *G.*-Group, *Gen.*-Gender, *P*-Participants, *Sess.*-Sessions, *f*-Female, *m*-Male, *nb*-Non-Binary, *D*-Dyads, *T*-Triads.

						Questionnaires			
G.	W1	W2	Gen.	P.	Sess.	12 it.	Self	Perc.	Cues
G1	R	C	f, f	2	10	20	20	20	20
G2	R	C	f, m	2	10	18*	18*	18*	18*
G3	R	C	f, f, nb	3	8	24	24	48	24
G4	C	R	m, m, m	3	10	30	30	60	30
G5	C	R	m, m	2	6	12	12	12	12
G6	C	R	f, f	2	10	20	20	20	20
Total	3-R	3-C	7-f 6-m	4-D	54	124	124	178	124
	3-C	3-R	1-nb	2-T					

3. Research questions

In this paper, we cover seven research questions (RQs) split into two sets. The first set of questions cover the effect of the avatar rendering style on communicative value, task satisfaction, and presence. The remaining set of research questions cover the ability to recognise others' mood. For ease of expression in the paper, we cover the two RQ sets separately in the Results (Section 5), but bring them together in the Discussion (Section 6).

RQs SET ONE

How do the avatar representations interact with:

RQ1: the functional communicative value based on (a) the identification of the other person (people); (b) the perceived authenticity of communications; (c) the perceived usefulness of expression and movement.

RQ2: the task satisfaction based on: (a) the level of task impact, (b) comfort and (c) engagement.

RQ3: the concept of presence based on: (a) co-presence and (b) social presence.

RQs SET TWO

RQ4: Does the avatar representation change the self-reported moods (a) overall and (b) over time?

RQ5: Does the avatar representation affect how accurately people perceive others' moods (a) overall and (b) do they improve overtime?

RQ6: Does positive or negative moods affect how accurately they are perceived by others?

RQ7: What are the most valuable cues available for identifying moods and are these different depending on the avatar rendering styles?

4. Methodology

4.1. Participants and tasks

Following ethics authorisation[1] we recruited participants in groups of 2 or 3 from the same company by sending out recruitment emails. The requirements for participation were that the individuals must know each other, work together, be part of daily work meetings, and be willing to conduct one of their regular daily meetings in mixed reality using HL2 for a period of 2 to 3 weeks (10 meetings). The number of meetings and the timeline was decided in line with previous longitudinal research (Han et al., 2022; Khojasteh and Won, 2021; Paulhus and Bruce, 1992). We offered a charity donation of 75.00 British Pounds per person on their behalf as an incentive. A total of 32 participants in 13 groups volunteered to take part, but 7 groups (18 participants) were unable to participate due to time and logistical constraints. As a result, a total of 14 participants (7 female, 6 male, 1 non-binary; aged 21–45) completed the study, forming 6 groups: 4 dyads and 2 triads. Out of these 6 groups, 4 were same-gender groups (2 male-only, 2 female-only), and 2 were mixed-gender groups. One of the 2 groups with 3 participants was a mixed-gender group, and the other was same-gender (see Table 1). All participants provided informed consent prior to taking part.

The members of each group remained the same throughout the study, and no participant missed a planned meeting. Some participants had the HL2 device at home (8 participants), while others were supplied with a device (6 participants) for the duration of the study. None of the participants had previously worked on remote MR meetings, although some had used the HL2 before. We installed the application on all of the HL2 devices. To maintain a high level of ecological validity, we did not ask the participants to perform a specific task. Instead, we allowed them to conduct their meeting as usual for at least 10–15 min. These meetings often took the form of daily stand-ups, status reports, or daily team catch-ups.

4.2. Avatars

Participants used full-body avatars in both a cartoon and a realistic rendering style. The avatar heads were personalised for each participant using a picture taken from the shoulders up. We used the local version of Avatar SDK (https://avatarsdk.com) to create the heads for both types of avatars (Cartoon: version 1.2.4; Realistic version 2.0.5). The heads were then attached to the bodies using Autodesk Maya (https://autodesk.co.uk/products/maya).

Both the Cartoon and Realistic bodies had the same skeleton structure and naming conventions, and there were four bodies available in total (two males and two females, one of each with a cartoon and realistic appearance). To minimise the impact of body variations, the bodies were very similar in appearance. Both types of avatars featured traditionally-proportioned human bodies wearing long trousers and long-sleeved polo neck sweaters, with the primary difference being that avatars had different coloured clothing (as shown in Figs. 1c-d and Fig. 2).

The avatars were animated in real time using inverse kinematics, with the input being the HL2 hand and head tracking signals. The hands moved when the HL2 detected hand movement using its external cameras, and the legs moved when the headset detected location movement based on the headset's position. Head pose (pitch, roll, and yaw) was animated based on IMU signals. Facial animation was generated using a lip-flapping script based on voice amplitude, with an additional periodic blinking animation. However, due to time constraints, the avatars did not have a sitting/standing animation or seated static position, so participants were instructed to stand for the duration of their meetings.

4.3. Device and application

The study was conducted using the Microsoft HL2 device (https://microsoft.com/en-gb/hololens). A networked MR application was developed using the Unity3D game engine [version *2020.3.12f1]* (https://unity.com) that allowed users to create, invite others to, and join remote meetings. In this application, users were represented as full-body avatars with two rendering styles of heads: cartoon and realistic. Users could see a hologram of a blue table (Figs. 1c,d) and a control menu (see supplementary materials). The table, which was adjustable and served as the centre of the meeting, was surrounded by the participants in the meeting. The control menu provided options for users to return to the *'Home'* menu, create a new meeting, see who is in the current meeting, join a meeting, mute themselves, adjust their microphone gain, switch their avatar, leave the meeting, and quit the application.

[1] Ethics authorisation was provided by Microsoft Research's Institutional Review Board (IORG0008066, IRB00009672)

Fig. 1. Example of people interacting in remote meetings using the HoloLens Mixed Reality device (a,b). Example meetings where participants are embodying realistic (c) and cartoon (d) full-body rendering style avatars.

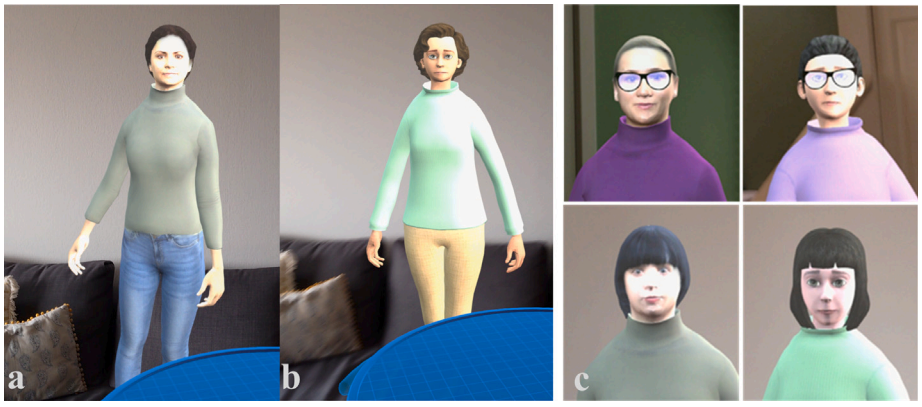

Fig. 2. Example of rendering styles of (a) realistic and (b) cartoon full body avatars in their partner local space with part of the adjustable blue table marking the centre of the shared space. (c): two examples of the same person embodying realistic (left) and cartoon (right) avatar rendering styles.

4.4. Procedure

After providing their consent, participants completed the demographic and on-boarding questionnaires and submitted a head and shoulders picture of themselves. This picture was used to create their cartoon and realistic avatars. The application was then installed on the HL2 devices and credentials were set up for each participant to access the application. Following this, the participating group and the researcher held a test meeting in MR to introduce the functionality of the application and perform a walk-through. The researcher was available to troubleshoot during each daily scheduled session, but not part of the meeting.

The procedure for each session was as follows: the participants, in their personal space (i.e., not in the lab), switched on the HL2 device, placed it on their head. Next, they opened the meeting application from the HL2 application menu, signed in with their credentials, and adjusted the blue table to ensure there was enough local space around it. This is because the rest of the group appeared around the table once the meeting was started. One group member created a meeting and added the other members to the meeting. The rest of the group joined the meeting as they were invited and changed their avatar to the corresponding one for that week (either Cartoon or Realistic). They then held their meeting as usual, after which they left the meeting and closed the HL2 application. The researcher was not present in this meeting, however, the researcher was always available to troubleshoot during the time the group's meeting was held. The only technical problem raised was with the HL2 application crashing. This was solved by restarting the application. This happened at the beginning of 7 different meetings (out of 54). Afterwards, the meeting continued, and hence, the data from these meetings was included in the analysis. Furthermore, the researcher could have joined the group meeting if this was needed. Following the meeting, the researcher reminded the group via text messages to complete the questionnaires for that session. This process was repeated until the final session.

4.5. Data

The participants completed consent forms and the following questionnaires: demographic, on-boarding (covering their expectations of having meetings in MR), and three daily questionnaire that they completed after each meeting.

Throughout the study, each group alternated between using one avatar type for half of their meetings and the other avatar type for the other half. In total 54 meetings were held, resulting in 124 daily questionnaire responses. However, one questionnaire was missing due to a technical issue during a meeting with cartoon avatars. This means there were 63 questionnaire responses from meetings with realistic avatars and 61 responses from meetings with cartoon avatars. Unfortunately, two groups were unable to complete all 10 sessions due to circumstances beyond our control: one triad had 8 sessions (with realistic avatars first) and one dyad had 6 sessions (with cartoon avatars first). Both triads also balanced the order of avatar use, with one starting with cartoon avatars and the other starting with realistic avatars (see Table 1). Half of the groups used first, in Week 1 (W1), Cartoon avatars, and then Realistic avatars in Week 2 (W2). The other half used Realistic avatars in W1 and Cartoon ones in W2. To avoid confusion about the starting avatar rendering style and the avatar rendering style used, we propose the following naming convention. We call the data from participants who used the Cartoon avatar in W1 followed by the Realistic avatar in W2 the *Cartoon–Realistic* condition (abbreviated as CR in figures). The data from participants who used the Realistic avatar in W1 and then used the Cartoon avatar in W2 was called the *Realistic–Cartoon* condition (abbreviated as RC in figures).

The daily questionnaire was divided into two parts. The first part contained 12 items with responses on a 1–7 Likert scale ranging from "strongly disagree" to "strongly agree". These items were selected and adapted from previous studies (Bailenson et al., 2003; Lombard et al., 2009; Harms and Biocca, 2004; Slater, 1999) to fit the design of the current study and to address RQs 1–3. This part focuses on communicative effects. For all questions see Table 2.

Table 2
The items in the daily questionnaire. Participants answered on a 1–7 Likert scale. *RQ* stands for *Research Question*. *m* and *sd* stand for mean and standard deviation, showing the descriptive statistics for each question. The * shows a statistically significant result of repeated Measures ANOVA comparing Cartoon and Realistic avatar rendering styles.

#	RQ	Questionnaire Item	Cartoon		Realistic	
			m	sd	m	sd
1	2c	I felt engaged in the meeting.	5.46	0.87	5.63	0.92
2	2c	I felt that my colleagues were engaged in the meeting.	5.41	0.95	5.55	1.04
3	1b	The avatars communicated like my colleagues.	3.6	1.57	3.9	1.41
4	2a	The appearance of the avatars affected the meeting tasks.	3.86	1.56	3.77	1.19
5	2b	The appearance of the avatars affected how comfortable I felt in the meeting.	4.05	1.6	3.86	1.55
6	3b	The appearance of the avatars mattered to me.	4.73	1.88	4.66	1.7
7	3a	I felt that I was in the presence of my colleagues.	4.67	1.49	5.18	1.6
8	1a	I could identify my colleagues.	5.12	1.54	5.78	0.98
9	3b	I perceive my colleagues' avatars as being only computerised images, not real people.	6.17	1.11	5.78	1.2
10	3b	There were obvious unnatural nonverbal behaviours from my colleagues' avatars.	5.34	1.27	5.48	1.23
11*	1b,c	The avatars' nonverbal behaviour was appropriate for the context.	3.08	1.36	3.79	1.04
12*	1c	The avatars' nonverbal behaviour was useful for understanding my colleagues.	2.72	1.15	3.55	1.26

The second part of the questionnaire focused on the recognition of moods and the usefulness of cues for perceiving these moods. Four moods were selected based on the UWIST mood checklist from Matthews et al. (1990): optimistic, focused, annoyed, and stressed. We define these moods as Matthews et al.: *[...] mood being defined here as an emotion-like experience lasting for at least several minutes. This definition distinguishes mood from cognitive evaluations per se, and from brief, phasic emotional responses to evaluations (Mayer, 1986)*. In the present study, participants were asked to rate their own moods and the perceived moods of their colleagues on a 1–7 Likert scale ranging from "strongly disagree" to "strongly agree". This was done at the end of the meeting for each moods covering the whole meeting. Participants in triads rated the perceived moods of the other two participants. The purpose of this questionnaire is to assess how accurately participants perceive moods based on the avatar rendering style. Hence, we include the self-reporting moods and the perceived moods of the participants' colleagues. Next, they were asked to rank 5 cues in order of their usefulness for recognising moods in their colleagues. These cues were: choice of words, movement/gesticulations, gaze, facial expressions, and tone of voice. The second questionnaire was used to address RQs 4–7.

4.6. Data analysis

Reverse coding.

After each meeting session, participants filled out two questionnaires, one on communication and one on moods. For the moods questionnaire, participants had to rate their moods on a scale from 1: *Strongly Disagree* to 7: *Strongly Agree*. They rated four moods: *Optimistic, Focused, Annoyed* and *Stressed*. Two of these had a positive connotation (Optimistic and Focused), while the other two had a negative connotation (Annoyed and Stressed). To calculate the overall self-reported rating of their moods, we utilised a reverse-coding technique for the negative moods (Annoyed and Stressed). This involved subtracting the ratings of these negative states from the maximum value (7: Strongly Agree) plus one ($7 + 1 = 8$). For instance, a rating of 5 for the mood Stressed would be transformed into a rating of 3 (8–5). These reverse-coded ratings were then utilised in our data analysis and to answer *RQ*4, as detailed in Section 5.2.

Accuracy of perceived moods.

We calculated the accuracy of the perceived mood by computing the error that participants had when perceiving the moods of their colleagues. We determined the error by mapping the absolute value of the difference between the self-reported mood and the perceived rating of the mood onto a scale of $[0, 1]$. With a maximum rating of 7 and a minimum rating of 1, the largest possible error was 6 $(7 - 1)$, which was mapped to a value of 1. The smallest error, which occurred when the self-reported rating was the same as the perceived rating of the mood, was mapped to a value of 0. For example, if the self-reported rating was 6 and the perceived rating was 2, then the error was 4 $(6 - 2)$; this was then mapped between $[0, 1]$, gaining the value of .667.

We calculated this error for each pair of participants in a group. In dyads, we considered the error for each participant in perceiving the mood of their colleague: P1's error in perceiving P2's mood ($P1\_to\_P2$) and P2's error in perceiving P1's mood ($P2\_to\_P1$). In triads, we considered each participant's error in perceiving the moods of all other participants in the triad. For example, in a group with participants P1, P2, and P3, we took into account all six possible combinations: $P1\_to\_P2$, $P1\_to\_P3$, $P2\_to\_P1$, $P2\_to\_P3$, $P3\_to\_P1$, and $P3\_to\_P2$.

We used this error to compare the overall error for Realistic and Cartoon avatars in RQ5a (Section 5.2). We also used it to address RQ5b on participants' ability to accurately perceive their colleagues' moods over time (Section 5.2). Finally, we used it to test RQ6 and the relationship between negative self-reported emotions and higher errors in perceived moods (Section 5.2).

5. Results

5.1. Communication, tasks and presence

Subsection Summary. In this subsection, we investigated how the avatar appearance interacts with the way participants communicate with each other, perceived task satisfaction and perceived sense of presence (RQ1-3). First, the participants perceived the realistic avatar's nonverbal behaviour as more appropriate for the interaction and more useful for understanding their co-workers compared to the cartoon avatar. Second, when looking at these responses over time, there were different insights for each avatar appearance based on which type the participants embodied first.

We first analysed the data from our within-group study by comparing the averaged scores for each participant using Cartoon and Realistic avatars. Next, we explored the effect of the passage of time on these scores by running regression models for each dependent variable and accounting for the temporal feature. A preliminary analysis of this data was presented in Dobre et al. (2022). In the current paper, we provide further work on this, covering the results in depth and putting them in relation to results from the moods analysis from Section 5.2.

Overview of the effect of realism

For each participant and for each question, we calculated two averages: one for all sessions (up to five) with the Cartoon avatar, and one for all sessions with the Realistic avatar. We then used a Repeated Measures ANOVA to assess the effect of realism on the data. The dataset did not violate the Repeated Measures ANOVA assumptions, including the Sphericity of the data. The descriptive statistics for this analysis can be found in Table 2, and a boxplot representation of the results for each question can be found in Fig. 3A.

RQ1: Functional communicative value. On average, participants reported higher scores for all four functional communicative value questions (Q3, 8, 11, 12). A Repeated Measure One-Way ANOVA found a significant difference for Q11 and Q12. The Q11 ($F(1,13) = 7.14, \mathbf{p = .019}, \eta^2 = .355$) shows that users rated the nonverbal behaviour of Realistic avatars more appropriate for the context (mean ± standard deviation: 3.79 ± 1.04) than the one of Cartoon avatars (3.08 ± 1.36), and Q12 ($F(1,13) = 5.5, \mathbf{p = .036}, \eta^2 = .296$) shows that users rated the nonverbal behaviour of Realistic avatars more useful for understanding their colleagues (3.55 ± 1.26) than the one of Cartoon avatars (2.72 ± 1.15). However, Q8 ($F(1,13) = 3.53, p = .08, \eta^2 = .217$) and Q3 ($F(1,13) = .718, p = .41, \eta^2 = .52$) were not significant, see Fig. 3A RQ1.

There was a significant interaction effect between avatar rendering style and order of use on participants' ratings of nonverbal behaviour appropriateness in Q11 ($F = 13.01, \mathbf{p = .004}, \eta^2 = .52$). This suggests that participants rated the Realistic avatars as more appropriate in terms of nonverbal behaviour, but only when they used the Realistic avatar first (Realistic W1: 3.9, Realistic W2: 3.7). On the other hand, the lower rating for Cartoon avatars was driven by those who used the Cartoon avatars first (Cartoon W1: 2.5, Cartoon W2: 3.6). These findings can be seen in Fig. 3A, as well as in Q11 and Q12. This result indicates that participants found their colleagues' nonverbal behaviour to be more appropriate for the context ($Q11$) and more useful for understanding their colleagues ($Q12$) when using the Realistic avatar rather than the Cartoon avatar.

RQ2: Task satisfaction. For task satisfaction, there were no significant differences between the two avatars in terms of the participants' level of engagement (Q1: $F(1,13) = .51, p = .49, \eta^2 = .04$), the perceived level of engagement of their colleagues (Q2: $F(1,13) = .44, p = .52, \eta^2 = .03$), the impact of appearance on the task (Q4: $F(1,13) = .08, p = .79, \eta^2 = .01$), or the reported level of comfort (Q5: $F(1,13) = .50, p = .50, \eta^2 = .04$).

RQ3: Presence. Once again, there were no significant differences between the two avatars in terms of the extent to which the avatar mattered to the participants (Q6: $F(1,13) = .07, p = .80, \eta^2 = .01$), the level of co-presence they felt (Q7: $F(1,13) = 2.1, p = .17, \eta^2 = .14$), or their perception of their colleagues' avatar as either digital images (Q9: $F(1,13) = 2.1, p = .17, \eta^2 = .14$) or unnatural (Q10: $F(1,13) = .44, p = .52, \eta^2 = .03$).

Overview of temporal effects

We were also interested in how participants' judgements of the avatars changed over time during the week in which they were using each avatar type. For this we calculated a linear regression between time (in days) and each of the questionnaire responses considered above. We computed the data for each avatar type, combining W1 and W2 (shown in Fig. 3B). We then present the data based on the avatar usage order (either Cartoon–Realistic (CR) or Realistic–Cartoon (RC)) in Fig. 3C and Fig. 3D.

RQ1: Functional communicative value. We found a significant positive correlation over time for being able to recognise their colleagues when participants embodied the Cartoon avatars ($R^2 = .06$, $F(1,59) = 4.25$, $\mathbf{p = .04}$), but not the Realistic avatars (shown in Fig. 3B for Q8). When separating the data by the order in which the avatars were used, the significance does not hold. The remaining questions for RQ1 (Q3, 11, and 12) did not show significance.

RQ2: Task satisfaction. When the order is not taken into account, there is no significance over time for task satisfaction (Q1, 2, 4, and 5; shown in Fig. 3B for RQ2). However, when considering the order, we see a significant decrease in participants' responses for the Cartoon avatars in the Cartoon–Realistic order ($R^2 = .13$, $F(1,29) = 4.18$, $\mathbf{p = .05}$, shown in Fig. 3C for Q5). This means that when embodying the Cartoon avatars first, their reported level of comfort was less influenced by the avatar's appearance over time. No other significant effects were found for Q5 or the other questions for RQ2 (Q1, 2, and 4).

RQ3: Presence. In terms of presence, the appearance of the avatar mattered less over time for participants using Cartoon avatars ($R^2 = .10$, $F(1,59) = 6.67$, $\mathbf{p = .01}$, Fig. 3B $Q6$). Additionally, participants using Realistic avatars reported fewer obvious unnatural nonverbal behaviours over time ($R^2 = .10$, $F(1,61) = 6.22$, $\mathbf{p = .01}$, Fig. 3B $Q10$). No other significant findings were discovered when examining data from both weeks (W1 and W2).

When examining the order in which avatar rendering styles were used, we found that the significance of Cartoon avatars in Q6 only remained for the Cartoon–Realistic order group when Cartoon avatars were used in the first week (W1: $R^2 = .27$, $F(1,29) = 11.06$, $\mathbf{p = .002}$, Fig. 3C $Q6$). When Cartoon avatars were used in the second week (Realistic–Cartoon order), there was a decrease but it was not significant (W2: $R^2 = .03$, $F(1,28) = .74$, $\mathbf{p = .39}$, Fig. 3D $Q6$). For the Cartoon–Realistic order group, there was also a significant drop for Realistic avatars in the second week (W2: $R^2 = .23$, $F(1,29) = 9.10$, $\mathbf{p = .005}$, Fig. 3C $Q6$), with the opposite trend observed for the Realistic–Cartoon order group in the first week, but it was not significant (W1: $R^2 = .04$, $F(1,30) = 1.38$, $\mathbf{p = .24}$, Fig. 3D $Q6$). Similarly, ratings of obvious unnatural nonverbal behaviours in the avatars showed that participants using Realistic avatars reported fewer of these over time during the Realistic–Cartoon order group in the first week (W1: $R^2 = .25$, $F(1,30) = 10.06$, $\mathbf{p = .003}$, Fig. 3D, $Q10$), but not during the Cartoon–Realistic order group (W2: $R^2 = .07$, $F(1,29) = 2.09$, $\mathbf{p = .16}$, Fig. 3C, $Q10$). No other significant findings were discovered for RQ3 on the other questions.

5.2. Perceived and self-reported moods

RQ4: Self-reported moods

Summary. In this subsection we investigated how the participants self-reported their moods when embodying the Cartoon or the Realistic avatars. When considering the overall data, we found (1) that participants reported more positive moods when embodying the Realistic avatars in the first week compared to the Cartoon avatars in the second week; and (2) that participants reported being more optimistic overall when embodying Realistic avatars. However, when we analyse the temporal data, we found that participants felt less optimistic and more stressed over time when embodying Realistic avatars in the first week.

RQ4a: Self-Reported Moods Overall. Participants self-reported their moods daily after each meeting. We were interested in the self-reported mood while participants embodied different avatar rendering styles and how this changes over time. We calculated the self-reported score for each avatar for each participant, averaging the data from Week 1 and Week 2. A paired t-test was conducted to compare the scores from Cartoon and Realistic avatars, regardless of the Order or Mood, and no significant difference was found (p= .98; mean± variance - C: 4.27 ± 1.28; R: 4.26 ± 1.26).

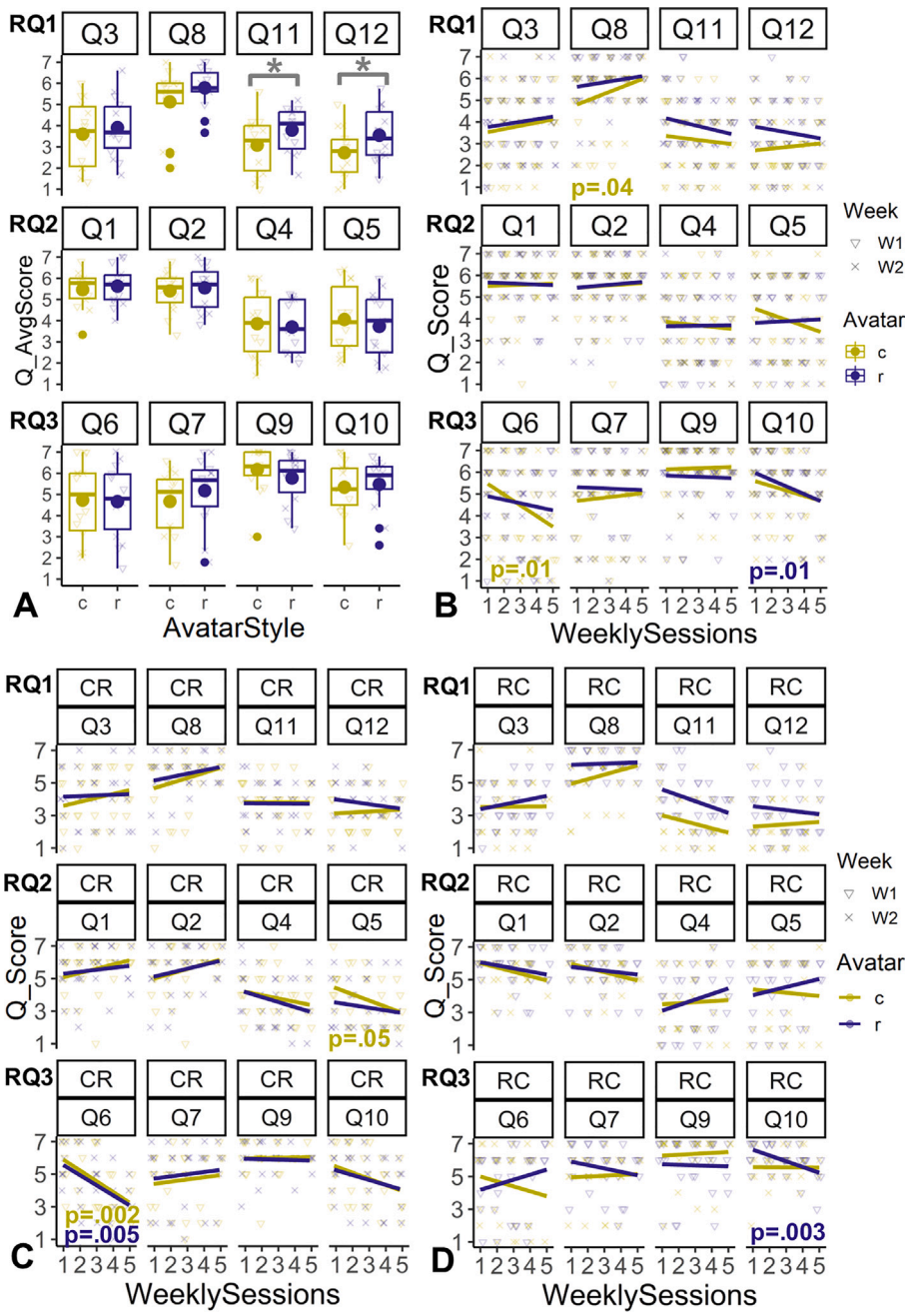

Fig. 3. Questionnaire responses from Table 2. Each row represents the questions for a RQ. Y-axis: averaged score (1=strongly disagree; 7=strongly agree). **A:** Boxplots per question separated by the avatar rendering style. **B-D:** Scatter plots showing each question score over time separated by the avatar rendering style. X-axis: weekly sessions in chronological order. In **C-D**, the question score is separated by the order (Cartoon–Realistic and Realistic–Cartoon).

As the order in which participants experience an avatar could influence their subjective experience, we also conducted separate paired t-tests between Realistic and Cartoon avatars for each condition. We found no significant difference in self-reported mood for the CR order ($t(27) = .45$, $p = .65$ C W1: 4.75 ± 2.07; R W2: 4.69 ± 1.83). However, for the RC order, participants reported more positive emotions when using Realistic avatars in Week 1 compared to Cartoon avatars in Week 2 ($t(27) = 2.81$, $\mathbf{p = .009}$, R W1: 4.9 ± 1.29; C W2: 4.23 ± 1.81), see Fig. 4(A). We also conducted paired t-tests for each Mood and for each order, but no significant results were found.

For each participant, we calculated the average score for each avatar and each Mood separately. Here we did not reverse the ratings for Annoyed and Stressed (i.e., a high in annoyed would indicate negative emotion). A paired t-test was conducted to compare the self-reported mood ratings for Cartoon and Realistic avatars, regardless of condition. Participants self-reported feeling more Optimistic in meetings using Realistic avatars compared to Cartoon ($t(13) = 2.53$, $\mathbf{p = .025}$; C: 4.6 ± 1.3; R: 5.06 ± 1.68), see Fig. 4B. There was no significant difference for the other emotions (Focused $t(13) = -.11$, $p = .92$, Annoyed $t(13) = -.57$, $p = .58$, or Stressed $t(13) = -.37$, $p = .71$).

RQ4b: Self-Reported Emotions overtime. The data was split by avatar rendering style and weekly session of avatar use, and regression analyses were conducted on the self-reported moods over time. We found a significant result for Realistic avatars used in the first week. Specifically, while using Realistic avatars in week 1, participants self-reported feeling less Optimistic ($R^2 = .14$, $F(1, 30) = 4.93$, $\boldsymbol{p = .034}$) and more Stressed over time ($R^2 = .16$, $F(1, 30) = 5.66$, $\boldsymbol{p = .024}$), see Fig. 4C. There were no significant results for the Focused and Annoyed moods

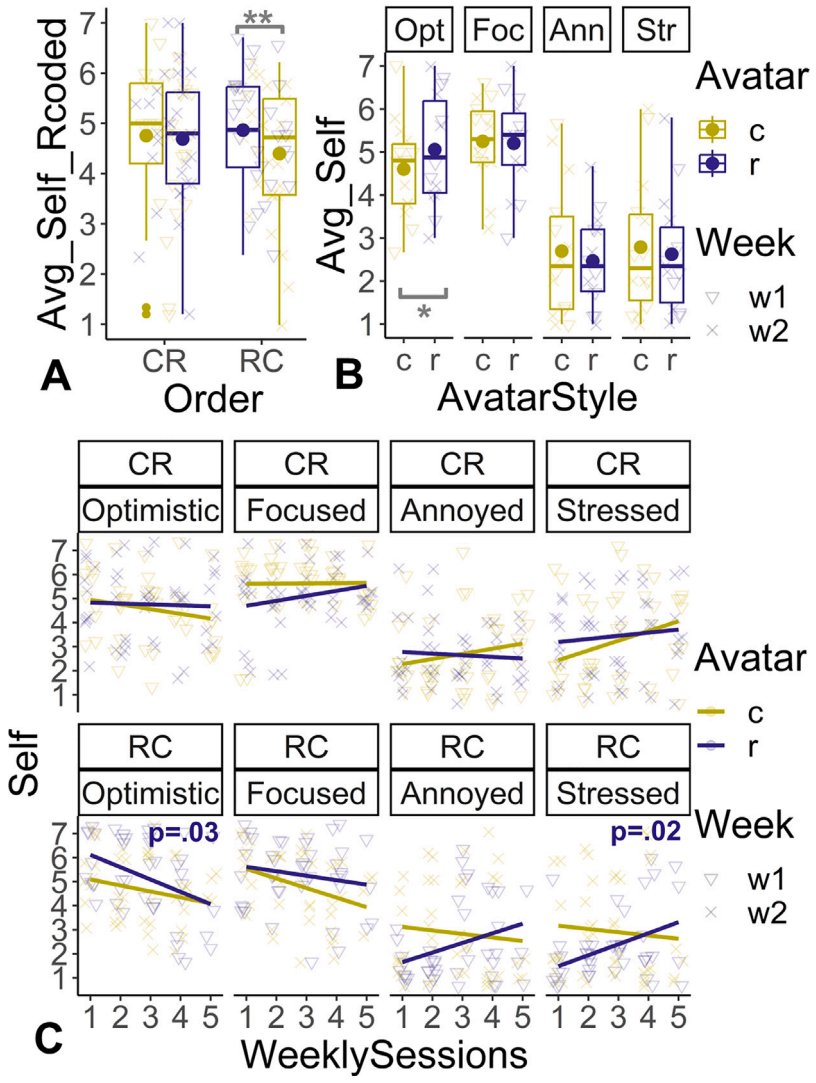

Fig. 4. Y-axis shows the self reported value: from 1 (Strongly Disagreeing) to 7 (Strongly Agreeing) of a certain mood. **A:** Average self-reported mood, with reverse coding for Stressed and Annoyed. The data is split by the order of Cartoon–Realistic and Realistic–Cartoon, and by the avatar rendering style. The X-axis shows the avatar order. **B:** Average self-reported ratings from each participant by mood and avatar rendering style. The self-reported ratings are separated by the avatar rendering style (Cartoon or Realistic) that participants were embodying. **C:** Self-reported ratings from each participant by order (Cartoon–Realistic in the top row and Realistic–Cartoon in the bottom row) and mood. The avatar rendering style is further represented by a fitted line. X-axis shows the weekly sessions chronologically from 1 to 5.

or for the Cartoon avatars. See supplementary materials for detailed statistics.

RQ5: Accuracy of perceived moods

Summary. In this subsection we investigated how accurately participants perceived their colleagues moods when embodying the two types of avatars rendering style. We found that, overall, there were higher errors when participants user Realistic avatars compared to Cartoon avatars. Higher errors came from participants embodied Realistic avatars in the first week compared to the second week, and from perceiving the negative moods (Annoyed and Stressed). Overtime, we found increasing errors for Annoyed and Stressed during the RC order. When separating by the avatar rendering style, participants embodying Realistic avatars had increasing errors for Optimistic and Annoyed during the first week and decreasing errors for Annoyed and Focused during the second week.

RQ5a: Accuracy of perceived moods overall. We were interested in the effect of avatar type on the accuracy of people's perception of their co-workers' moods. As we performed a Repeated Measures ANOVA test, we checked that the dataset did not violate the measure's assumptions, including the Sphericity of the data. Hence, a repeated measure 2 × 4 ANOVA on the Normalised Error using avatar rendering style (Realistic, Cartoon) and mood (Optimistic, Focused, Annoyed, Stressed) as within-subjects factors, and order (CR and RC) as a between-subjects factor showed a significant difference in the error of perceived moods when using Cartoon versus Realistic avatars ($F(1, 18) = 5.13$, $\mathbf{p = .036}$, $\eta^2 = .22$). Specifically, participants perceived their colleagues' moods with fewer errors when using Cartoon avatars (M = .22) compared to Realistic ones (M = .25).

We found a significant interaction effect between the order and the type of avatar ($F(1, 18) = 5.91$, $\mathbf{p = .026}$, $\eta^2 = .247$). This is evident of more errors for Realistic avatars when they were used in the RC order compared to CR. Specifically, when participants used Realistic avatars in week 1, their mean error rate was .270, but .226 when Realistic avatars were used in week 2.

For the interaction effect between avatar rendering style and order, we conducted a post-hoc analysis using a paired t-test to compare errors made by participants in the CR and RC orders. No significant difference was found in the CR order ($t(9) = .11$, $p = .91$; C W1: .228 ± .006; R

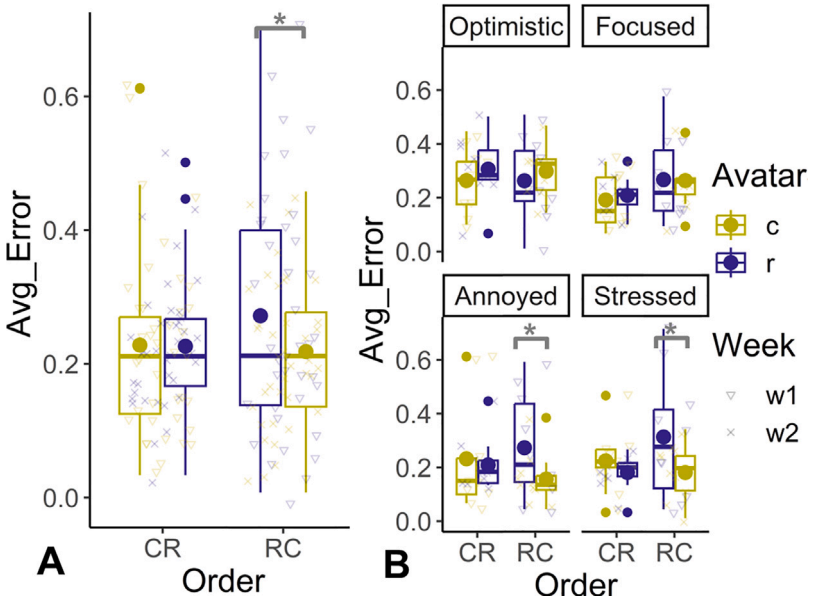

Fig. 5. Boxplots showing the averaged mapping error of the perceived mood of each participant to that of their colleagues. The averaged mapping error on the Y-axis ranges from 0 (no error) to 1 (maximum error possible). **Panel A:** The error values are separated by order (Cartoon–Realistic or Realistic–Cartoon), and the data from each order is further separated by avatar rendering style (Cartoon or Realistic). The shape of each data point indicates the week (W1 or W2) in which the data was collected. **Panel B:** The data from each avatar rendering style is separated by order and by mood, with each boxplot showing the error of perceiving a specific mood (Optimistic, Focused, Annoyed, or Stressed), separated by avatar rendering style.

W2: .226 ± .003). However, in the RC order, participants made more errors in perceiving their colleagues' moods while using the Realistic avatar in the first week and then the Cartoon avatar in the second week ($t(9) = 3.79$, $\mathbf{p = .004}$; R W1: .27 ± .004; C W2: .21 ± .001), see Fig. 5A.

There was also an interaction effect between the order, the avatar rendering style, and the mood of the participant ($F(3, 18) = 3.71$, $\mathbf{p = .017}$, $\eta^2 = .171$), see Fig. 5B. When examining the data by mood, we found that the error rates varied depending on the order and rendering style of avatar used. For Cartoon avatars, the error rate was generally higher for the positive moods (Optimistic and Focused) in the RC order compared to the CR order. However, for the negative moods (Annoyed and Stressed), the pattern was reversed, with the error rate being higher for the CR order. For Realistic avatars, the error rate was higher for almost all moods except for the Optimistic state.

A two-factors ANOVA with avatar rendering style (Cartoon and Realistic) as the dependent variable and order as the between-subjects factor revealed significant differences between Cartoon and Realistic avatars for the Annoyed ($F = 4.41$, $\mathbf{p = .05}$) and Stressed (F=5.32, $\mathbf{p = .033}$) moods, but no significant differences were found for the Optimistic (F=2.77, $p = .11$) or Focused ($F = .11$, $p = .74$) moods.

RQ5b: Accuracy of perceived moods over time. We were interested in checking whether people got better at perceiving their colleagues' moods over time. To test this, we used the normalised error of the perceived mood as presented in Section 5.2. As we also considered the time variable, in this case we did not average the error over time as in RQ5a. First, we considered all data regardless of which avatar the participants were using, then we separated this data based on the Order (Cartoon–Realistic or Realistic–Cartoon), and finally we presented the results for Cartoon and Realistic avatars.

We calculated the regression statistics over time for each mood for all 10 sessions over the two weeks, maintaining chronological order. There was no significant trend for any of the moods (Optimistic: $R^2 = .007, F(1, 178) = 1.21$, $p = .27$, Focused: $R^2 = .002, F(1, 178) = 0.44$, $p = .51$, Annoyed: $R^2 = .007, F(1, 178) = 1.29$, $p = .26$, Stressed: $R^2 = .007, F(1, 178) = 1.2$, $p = .27$). We also calculated the regression over all 10 sessions separated by order. There was no significant result for the Cartoon–Realistic order for any of the moods (Optimistic: $R^2 = .004, F(1, 90) = .37$, $p = .54$, Focused: $R^2 = .01, F(1, 90) = 1.23$, $p = .27$, Annoyed: $R^2 = .008, F(1, 90) = .7$, $p = .40$, Stressed: $R^2 = .01, F(1, 90) = 1.18$, $p = .28$). However, for the Realistic–Cartoon order, there was a significant increase in error over time for the moods Annoyed ($R^2 = .06, F(1, 86) = 5.46$, $\mathbf{p = .022}$) and Stressed ($R^2 = .05, F(1, 86) = 4.88$, $\mathbf{p = .029}$). There was no significance for Optimistic ($R^2 = .001, F(1, 86) = .86$, $p = .36$) or Focused ($R^2 < .00, F(1, 86) = .01$, $p = .92$). Fig. 6A shows the trends for the Cartoon–Realistic and Realistic–Cartoon orders.

Next, we separated the data by the avatar rendering style. We calculated the regression for each avatar rendering style considering the order they were used in the weekly meetings, each of which had data from 5 meetings (see Fig. 6B). We found significant results for the Realistic avatar only. For those who started using Realistic avatars in their first week (W1), there was an increase in the error over time of perceiving others as Optimistic ($R^2 = .086, F(1, 42) = 3.96$, $\mathbf{p = .05}$) and Annoyed ($R^2 = .11, F(1, 42) = 5.48$, $\mathbf{p = .02}$). However, for those participants who did not use Realistic avatars until their second week (W2) (CR condition), there was a decrease in the error for Annoyed ($R^2 = .13, F(1, 44) = 6.43$, $\mathbf{p = .01}$) and Focused ($R^2 = .3, F(1, 44) = 19.91$, $\mathbf{p < .001}$). Results with non-significant p-values can be found in the supplementary material.

We conducted a repeated measure 2 × 4 × 4 ANOVA with avatar rendering style (Realistic, Cartoon), mood (optimistic, focused, annoyed, stressed), and time (T1-T4). We only included the first four meetings with each avatar rendering style in each week due to missing data on the fifth day (some groups of participants only completed four meetings with each avatar rendering style). We also removed data from one group of two participants that only completed three sessions with each avatar rendering style. Therefore, for this analysis we used data from 12 participants over four sessions with each avatar rendering style (a total of eight sessions). We checked that the dataset did not violate the Repeated Measures ANOVA's assumptions, including the Sphericity of the data.

We found a significant result for avatar rendering style ($F = 8.44$, $\mathbf{p = .01}$, $\eta^2 = .33$), with participants using realistic avatars having higher errors ($M = .247$) than those using cartoon avatars ($M = .207$). This result is consistent with the overall result from RQ4 in Section 5 and Fig. 5A, which showed that participants had higher errors with realistic

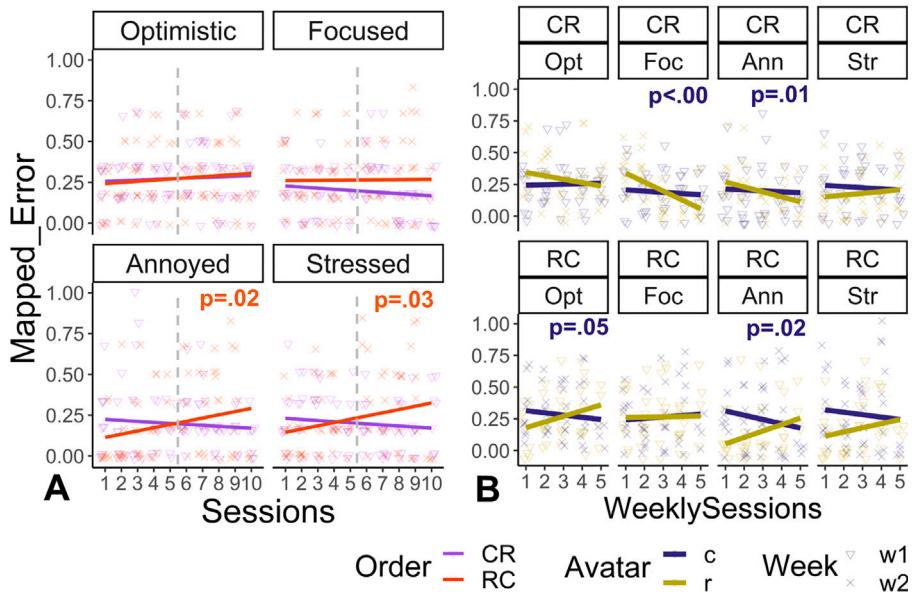

Fig. 6. Scatter plots of the average mapping error of perceived mood from each participant to each of their colleagues. The mapping error on the Y-axis ranges from 0 (no error) to 1 (maximum error possible). **Panel A**: The X-axis represents the order each group had chronologically, with W1 followed by W2, in Cartoon–Realistic order or Realistic–Cartoon order. The grey dotted lines separate the plots into weekly groups, with the first part being W1 and the second W2. Additionally, the errors were separated into different sub-graphs by mood, from left to right: Optimistic, Focused, Annoyed, and Stressed. **B**: The X-axis represents the sessions each group had during one week (1 to 5). The data is separated into weekly groups: week 1 (W1) is shown in the top graphs and week 2 (W2) is shown in the bottom graphs. The moods were also separated from left to right: Optimistic, Focused, Annoyed, Stressed. Additionally, trend lines are fitted for each week and mood to show the errors for each avatar rendering style (Cartoon or Realistic).

avatars compared to cartoon avatars, even when not considering the fifth session with each avatar rendering style that had fewer data points.

We also found an interaction effect on the mood and time (F=2.34, $\mathbf{p = .017}$, η^2 = .12) and on the avatar rendering style, mood and time (F=2.83, $\mathbf{p = .004}$, η^2 = .14).

The participants starting with Cartoon avatars had a significant lower error (M = .20) compared to those starting with Realistic avatars (M = .25), hence the starting avatars having a significant between-subjects effect (F = 5.7, $\mathbf{p = .03}$, η^2 = .26).

RQ6: Self-rated moods and co-workers' perception errors

Summary. In this subsection we investigated the self-reported moods and the link to the co-workers' perceived error of these moods. For the negative moods, the results show a high error for when participants self-report high levels of these moods, and low errors for low self-reported levels of Stressed (only CR order) and Annoyed (both CR and RC orders). This trend is reversed for Focused during the CR order: self-reported high levels have a low error rate, and self-reported low levels have a high error rate.

Khojasteh and Won (2021) demonstrated a link between high levels of self-reported positive emotions and increased accuracy in perceiving those emotions by others. They also found an opposite trend for negative emotions, with high levels of self-reported negative emotions leading to decreased accuracy in perceiving those emotions. We sought to replicate these results and found similar outcomes. We observed similar correlations for Focused, Stressed, and Annoyed, but no correlation for Optimistic. To compute these results, we used the error of the perceived mood and the self-reported ratings of participants' moods. We conducted regression analyses for each avatar and for each week (Fig. 7).

When participants self-reported high ratings of Focused, their colleagues had fewer errors in perceiving it. Conversely, when participants self-reported low ratings of Focus, their colleagues showed more errors in perceiving it. This trend was only significant for Realistic avatars in the second week (R^2 = .15, $F(1, 44)$ = 7.99, $\mathbf{p = .007}$). No significance was found for Cartoon avatars.

The trend was the opposite for Annoyed, and it was significant for both Cartoon and Realistic avatars during the first and second week. Specifically, when participants self-reported high ratings of Annoyance, their colleagues showed more errors perceiving it. When the self-reported ratings were low, their colleagues' perception of that emotion had fewer errors. This result was significant for Cartoon avatars (W1: R^2 = .15, $F(1, 44) = 8.11$, $\mathbf{p = .006}$, W2: R^2 = .09, $F(1, 42) = 4.6$, $\mathbf{p = .03}$) and Realistic avatars (W1: R^2 = .24, $F(1, 42)$ = 13.53, $\mathbf{p < .001}$, W2: R^2 = .39, $F(1, 44) = 28.8$, $\mathbf{p < .001}$).

Stressed had the same trend as Annoyed, but the results were only significant for the Realistic–Cartoon order (Realistic W1: R^2 = .12, $F(1, 42)$ = 5.84, $\mathbf{p = .02}$, Cartoon W2: R^2 = .14, $F(1, 42)$ = 7.02, $\mathbf{p = .01}$).

RQ7: Most useful emotional cues

Summary. In this subsection we investigated the participants' most useful cues. We found a difference only for the facial expression. When participants started with Cartoon avatars, they rated the facial expressions cue more useful than when they started with Realistic avatars (CR order vs RC order).

Participants were asked to rank the usefulness of various cues in perceiving the moods of their colleagues. On average, the cue of *choice of words* was rated the highest, followed by *tone of voice* and *movements/gesticulations*, for both Realistic and Cartoon avatars, and in both orders. The cues of *gaze* and *facial expression* were rated the lowest. In the Cartoon–Realistic order, both *gaze* and *facial expression* had a similar average score (both were equally the least useful). However, in the Realistic–Cartoon order, *gaze* was rated more useful than *facial expression* (see Fig. 8).

We compared the data from the first and second weeks for both Cartoon and Realistic avatars based on the Order. We found a significant difference for the cue of *facial expression*. For Cartoon avatars in the first week, participants rated facial expression as more useful than in the second week ($t(18) = -3.8, p = .001$; Cartoon W1: mean = 4.37,

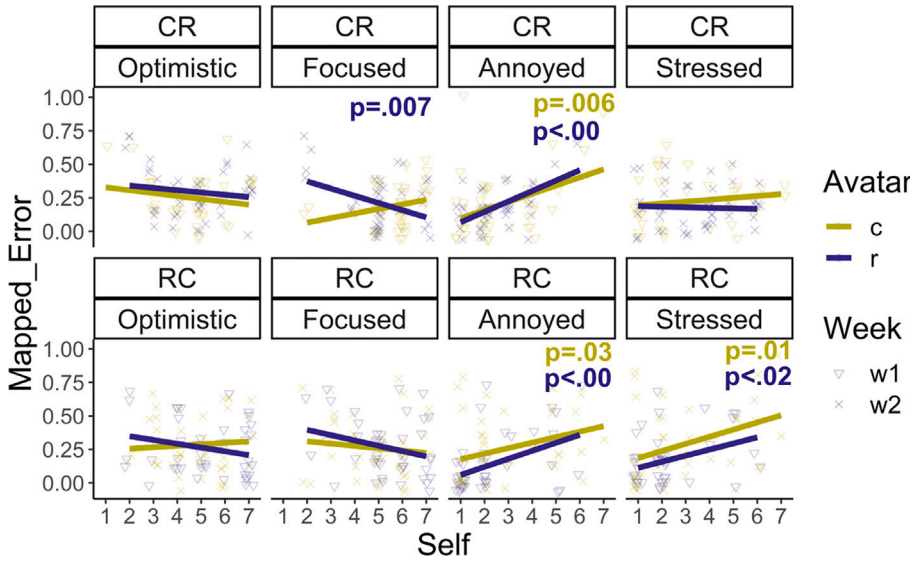

Fig. 7. Scatter plots of mapped error of perceived mood from each participant to each of their colleagues. The mapped error on the Y-axis ranges from 0 (no error) to 1 (maximum error possible). The X-axis represents the self-reported mood, 1 representing *Strongly Disagree* and 7 representing *Strongly Agree* with having a certain mood. The data is separated first by the order (Cartoon–Realistic (top row) and Realistic–Cartoon (bottom row)) and then split by each mood (Optimistic, Focused, Annoyed, Stressed), and the data points are fitted for each avatar rendering style (Cartoon and Realistic).

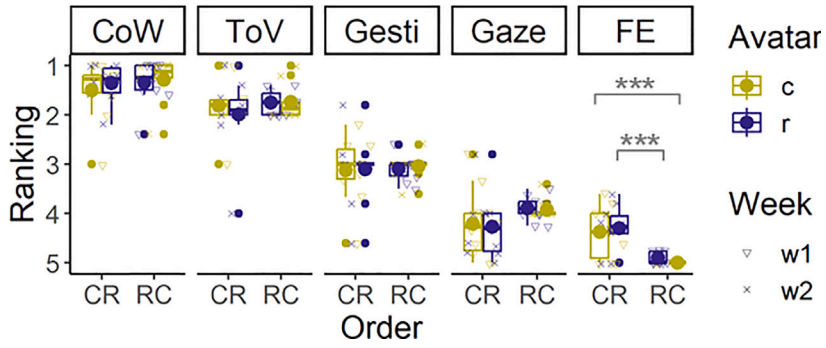

Fig. 8. Box plots of the ranked cues used to perceive the other's moods. On the Y-axis, there is a 1 – 5 ranking, with 1 meaning the most useful cue and 5 the least useful.

variance = 0.27; Cartoon W2: mean = 5, variance = 0). Using Realistic avatars in the first week, the rating of facial expressions was lower than in the second week ($t(18) = 4.11, p < .001$; Realistic W1: mean = 4.9, variance = .01; Realistic W2: mean = 4.3, variance = .2).

6. Discussion

The appearance of avatars in social interactions seems to have various complex implications, some of which depend on the order in which the two avatars are used (i.e. Cartoon–Realistic or Realistic–Cartoon). In the remainder of this section, we discuss three topics: the impact of high expectations when using Realistic avatars first; the process of becoming accustomed to the Cartoon rendering style of the avatars; and finally, the greater occurrence of errors in perceiving colleagues' negative emotions. We then comment on the implications of these findings for the design and deployment of avatars for MR meetings.

Realistic avatars may lead to high expectations

When participants embodied Realistic avatars, they perceived their colleagues' nonverbal behaviours as being more useful and appropriate for the interaction compared to when they used Cartoon avatars (RQ1b $p = .019$, RQ1c $p = .036$). There was also an order effect for appropriateness, showing that the high scores primarily came from when participants used Realistic avatars in Week 1, rather than in Week 2 ($p = .004$). This result suggests that participants may have had higher expectations when using Realistic avatars in meetings with their colleagues. While there was no difference in the nonverbal behaviour of Cartoon and Realistic avatars, participants still rated the nonverbal behaviour of Realistic avatars as having more functional communicative value (RQ1).

On average, participants rated their feelings of being in the presence of their colleagues higher for Realistic than for Cartoon avatars, although this difference was not significant ($p = .17$). When embodied in Realistic avatars, being able to clearly identify their work-colleagues right from the start of the study, as in real meetings, may have contributed to a higher sense of presence.

Participants self-reported feeling more Optimistic when using Realistic compared to Cartoon avatars (RQ4 $p = .025$). However, participants self-reported feeling more Stressed and less Optimistic over time when they used Realistic avatars in Week 1 of the Realistic–Cartoon order (RQ4 more Stressed $p = .024$, less Optimistic $p = .034$). Work meetings have their own stresses, so it may be that Realistic avatars conveyed those stresses, or it may be the case that as participants found Realistic avatars not to live up to initial expectations, stress increased and optimism decreased. Given the lack of a pre-meeting, baseline mood questionnaire, we might inferred this from participants'

comments such as: *'Overall I was really impressed with the likeness of the avatars to the real people'*, reported after the first time using the realistic avatars in session 1 in R-W1. And the comment: *'The voice part is ok, however the avatar does not match the expression of my colleagues. Having the ability to work with them face to face, I can easily tell when there are issues.'*, from the last session in R-W1. This shows how the initial expectations did not last. However, other confounding variable could have impacted this result. Hence, further research is needed to separate the mood effects of avatar appearance from the mood effects of the work/task being done.

The potential higher expectations of realistic avatars may have led participants to rate gaze cues as being more useful than facial expressions for perceiving their colleagues' moods (RQ7 $p = .001$). The avatars had neither true gaze cues from eye-tracking nor facial expressions from face-tracking or audio visemes, but they did have head pose relative to body pose and a blink animation. Participants in the Realistic–Cartoon order condition considered gaze to be more useful than facial expression, which may be due to their higher expectations when using Realistic avatars first, especially since heads could turn and eyes blinked, which may have conveyed an illusion of gaze.

Finally, participants embodying Realistic avatars first (in the Realistic–Cartoon order) had more errors when perceiving their colleagues' moods (RQ5a). There was an increase in errors over time for the Realistic–Cartoon order for perceiving Annoyed and Stressed emotions (RQ5b $p = .022$ and $p = .029$), as well as increased errors in perceiving Optimism and Annoyed emotions (RQ5b $p = .05$ and $p = .02$) for those who used Realistic avatars in Week 1. Nonverbal behaviour was implemented in the same way for both Cartoon and Realistic avatars. However, the higher expectations of avatars resembling their colleagues in a Realistic manner may have led to an assumption that they provided authentic and more useful expression and movement (RQ1 (b) and (c)). Given that participants in the Realistic–Cartoon order also reported gaze behaviour as more useful than those in the Cartoon–Realistic order (RQ7), the combination of effects could have led to more errors in the perception of moods.

In summary, the Realistic avatars may have led participants to have unrealistic expectations of the accuracy of the body language presented by those avatars, despite the fact that the body language was identical in both Realistic and Cartoon avatar rendering styles. Errors in reading other participants' emotions may have occurred because participants put too much trust in non-verbal cues relative to the more accurate verbal cues. On the other hand, participants' recognised their colleagues' identities more easily with Realistic faces, this could explain why Realistic avatars provided a greater sense of presence than Cartoon avatars.

Participants may become accustomed to cartoon avatars over time

When using Realistic avatars, participants consistently rated their ability to identify colleagues as a 5.8 on a scale from 1 to 7. However, when using Cartoon avatars, participants showed an improved ability to identify their colleagues over time (RQ1a, $p = .04$). These results held even though both the Cartoon and Realistic avatars were personalised for each participant using a picture of themselves (see Section 4).

At the same time, participants reported that the appearance of the avatar mattered less to them over time, when taking part in the Cartoon–Realistic order condition. When using Cartoon avatars in their first week, their score on the question 'The appearance of the avatars mattered to me' dropped from a 5.7 (agree) to a 3.2 (slightly disagree) (RQ3b, $p = .002$). A similar trend was observed when they then switched to using Realistic avatars. On the first day using the Realistic avatars in their second week, the average score to Q6 was 5.6 (agree), falling to 3.2 (slightly disagree) by the fifth day (RQ3b $p = .005$). Additionally, participants reported feeling more comfortable using Cartoon avatars over time in the first week (RQ2b, $p = .005$). These trends were not observed when participants used Realistic avatars before Cartoon avatars (Realistic–Cartoon order) or for Realistic avatars alone. It is possible that using Cartoon avatars first allowed participants to become more accustomed to the appearance of the avatars, leading to increased comfort and a reduced focus on appearance. This may also have contributed to the improved ability to identify colleagues and the decreased reliance on facial expression and gaze as cues for perceiving moods (RQ7, $p < .001$).

Overall, participants using Cartoon avatars made fewer errors in perceiving the moods of their colleagues compared to those using Realistic avatars (RQ5a, $p = .036$). This trend was particularly pronounced in the first week of the study when using Realistic avatars (RQ5a). In contrast, errors decreased over time for both Focused and Annoyed moods when using Realistic avatars in the second week (RQ5a, $p < .001$ and $p = .01$, respectively). However, the use of Realistic avatars in the first week was associated with an increase in errors for both Optimistic and Annoyed moods (RQ5a, $p = .02$ and $p = .05$, respectively). These findings might imply that using the Cartoon avatars first did not lead to as high expectations as might have happened in Realistic–Cartoon order, with a subsequent sense that appearance mattered less over time, and leading them to rank facial expression and gaze as equally not useful for perceiving moods (RQ7). Further, a decreased emphasis on visual appearance may have led participants in the Cartoon–Realistic order condition to focus more on the less-mediated auditory cues for moods, potentially leading to greater accuracy. Additionally, avatars were not implemented with gaze and facial expressions, which could have affected these results. At the same time, the auditory cues were also the ones people relied the most on in two other prior longitudinal works (Khojasteh and Won, 2021; Moustafa and Steed, 2018).

In summary, time had an important impact on how avatar appearance influenced participants, supporting the rationale for and further need for longitudinal studies. The appearance of the avatar mattered less to participants over time. This could be for several reasons. Firstly, it could simply be that participants grew accustomed to the different avatar rendering styles over time and were able to use them effectively. Secondly, given that the two avatar types did not have functionally different behaviour, the difference made little practical difference beyond an initial novelty effect. Finally, the main functional difference between avatars seems to be in the ability to recognise other participants, which matters less over time as participants get used to their colleagues' avatars. This has been reported also in other longitudinal studies where participants were getting used to the environment and the avatars (Khojasteh and Won, 2021; Bailenson and Yee, 2006; Moustafa and Steed, 2018).

Use of avatars may lead to more errors perceiving negative emotions

Participants made more errors when trying to perceive their colleagues' negative moods. This finding is consistent with the result of Khojasteh and Won (2021). However, this trend was not observed for all moods. For Annoyed, this trend was observed for both the Cartoon and Realistic avatars, regardless of the order condition (Cartoon–Realistic or Realistic–Cartoon). For Stressed, participants made more errors in perceiving negative emotions only when in the Realistic–Cartoon order. The opposite applied to Focused, in which there were more errors during the Cartoon–Realistic order. There was no trend for errors in perceiving the Optimistic mood.

This result might explain some of the outcomes for errors of different avatar rendering styles, for instance, the higher error when perceiving colleagues' emotions when they used Realistic avatars. These errors might come from participants self-reporting more negative emotions, hence colleagues making more errors when trying to perceive those emotions. However, while participants self-reported as more optimistic in Realistic avatars compared to Cartoon avatars (see Section 5.2), this did not result in colleagues making more errors in perceiving optimism, as they did for negative emotions. There was no significant difference between self-reported moods when using Realistic

and Cartoon avatars overall that would explain the higher error rate for perceiving negative emotions in Realistic avatars.

Most of the error in perceiving moods occurred in Realistic W1, so we also compared the self-reported mood for each avatar and week. As shown in Section 5.2, there was no significant difference between Realistic W1 and Cartoon W1, regardless of the mood or when each mood was considered separately. There was a difference in the self-reported moods between avatar rendering styles in the Realistic–Cartoon order. Participants in Realistic avatars in their first week self-reported their moods more positively than they did when using Cartoon avatars in their second week (RQ4 p= .009). This contradicts the implication that the higher error when perceiving emotions while embodying Realistic avatars (RQ5a) might be a result of more negative moods.

In summary, errors in recognising emotions via avatars were not uniform, with greater errors made for negative emotions. This may be due to people tending to mask negative emotions more (and masking being easier with avatars that show limited non-verbal cues). Errors were greater when using Realistic avatars before Cartoon ones (with the exception of Focused), perhaps because of the greater expectations people had of realistic avatars (as discussed above). At the same time, Suma et al. (2023) showed that the avatar's voice and facial expression could affect the capabilities to recognise emotions .

Implications

First Impressions. The findings above highlight a key issue: in today's early stages of mixed reality meetings in the workplace, the first impressions of avatars set expectations. Both realistic and cartoon avatars are useful for establishing co-presence. However, with the current, limited abilities of avatars to visually convey the *nuance of communication or emotion*, the communicative value of realistically rendered avatars is more fragile than that of cartoon avatars, leading people to misplaced belief in their non-verbal behaviour such as gaze and emotions.

Importance of technical maturity of the system. At this point in both research and commercial history of avatars, these findings are probably indicative of an immature system of realistic avatar production and their relative novelty to users. While there will be occasional potholes on the way up the recovery curve of the uncanny valley, given current advancements in methods for creating and animating realism, realistic avatars are on fast track to acceptability (Khakhulin et al., 2022; Zhang et al., 2022; Ma et al., 2021). The question for commercial systems will be how to set users' expectations for avatar use. One way could be to deliberately separate avatar rendering styles by context (e.g., cartoon for casual and realistic for business, as Mark Zuckerberg believes (Fridman, 2022)). The results of this study suggest that having clear evaluations of how well avatars enable identification of a unique individual, communicative functionality, and emotional trustworthiness may matter more than thresholds of accuracy in realistic depiction. In particular, our participants made a range of errors in perceiving positive and negative moods, and these errors were different for cartoon versus realistic avatars. These results show that it will be important to disentangle issues of accuracy of likeness from issues of emotional trustworthiness of likeness.

Importance of Verbal Cues. This brings us to a related point about what participants were basing their perceptions on. Although the object of primary comparison in this study is differences in visual representation, it is also crucial to note how important verbal cues were to participants. In the cartoon version, participants were highly attentive to verbal cues. In traditional video meetings, voice is often the communicative stream of primary value, especially for some of the most common business needs (Standaert et al., 2021). This has two implications. First, in the short term, improving audio quality (e.g., designing even better spatial audio, which is already quite good in VR and MR systems) may provide more impact than improving visual realism. Although recent studies by Immohr et al. (2023, 2024), Fink et al. (2024), showed mixed results when it comes to improvements of social interaction while using spatial and non-spatial audio in dyadic settings, future work should further analyse this including in larger groups. Second, designing better cartoon *and* realistic avatars should involve detailed consideration of the interaction between the visual and the verbal. Specifically, new avatars must not be evaluated as still images only or silent video. Their value will come as a holistic system, and it is in that holism that useful trade-offs in visual realism will be found.

Importance of evaluation over time. Finally, in terms of holistic understanding, time also matters. Our results suggest that even fairly short longitudinal studies, e.g., daily use for around two weeks, produce important changes in how people perceive avatars, especially when those people know one another. More research will be needed to determine how many instances of multiple exposures to different styles might be needed, over what time period, and to what extent of acquaintance will provide the strongest results. We hasten to add that we are not claiming that all short-duration research between strangers is problematic or has no ecological validity. There are many situations of short-duration communication between strangers in work contexts, ranging from fleeting transactional encounters (Félix-Brasdefer, 2015) to v-teams who come together under conditions of swift trust (Meyerson et al., 1996; Blomqvist and Cook, 2018). We would urge that future research on avatars features multiple encounters over time, lest we over-index on first impressions instead of allowing that time will tell.

7. Limitations

Although we balanced the participants' gender, group size and avatar order, we did not take into account the participants' prior experience with MR devices. Six out of 14 of them never used a MR device before, whereas the rest had some experience with it in the past six months (two participants – more than two times a week, two participants – once a week, four participants- 1–3 times a month). Given that this study has a relatively small number of participants, we could not test if the prior MR experience influenced participants' responses. Thus, another limitation is the relatively small number of participants. Further studies should look at larger sample sizes to investigate the effect of prior MR experience and the interaction effects between different factors.

Additionally, the core contribution lies in the methodology of this study. We conducted a high ecological validity study where repeated real-world meetings were held in mixed reality for 2–3 weeks, using two different types of personalised avatars. Given the novel methodology, the dataset has some limitations. First, this impacted the number of participants. Furthermore, as the data was collected during meetings between groups of co-workers, participants' questionnaires responses could be influenced by the interaction within the work meeting. Future work should investigate similar settings to confirm our results.

All participants were part of a technology company. Hence, there might be a possibility that they were more accepting of innovations in IVEs. Further studies are needed to control for the likelihood that participants might be more open to novel technology systems.

Due to limited time, we did not implement into the avatars non-verbal behaviour gaze or facial expressions. In particular, the results on RQ7 (*What are the most valuable cues available for identifying moods and are these different depending on the avatar rendering style*) might have been different if more detailed gaze and facial cues had been implemented. Additionally, participants were embodying avatars with similar simple clothing style, but with different coloured tops (see Fig. 1e). Although participants did not comment on the clothing style or colour, they might have affected the way they perceived their co-workers.

Although we designed the study to answer the set research questions, we assigned 1–2 questionnaire items for each research question. We took this decision to avoid a very lengthy questionnaire. This has

also led to rich but complex and dense results. Future studies are needed for further and more in-depth analysis of a similar longitudinal field study design.

We consider the participants' self-reported emotional states over time for RQ4b (*Does the avatar representation change the self-reported moods over time?*), however, we did not compare these to a pre-meeting baseline rating. Because of this, our results regarding the self-reported moods could be influenced by external factors. This research question should be explored in more detail in future studies that include pre and post questionaries.

8. Conclusion

We presented the results from a longitudinal study on avatars' appearance during work-related meetings between co-workers. We investigated how the avatar appearance interacts with: the way participants communicate with each other, perceived task satisfaction, perceived sense of presence, emotional state perception, and useful cues in MR meetings. Over two-three weeks, 14 participants in dyads and triads (6 groups) had their usual work meetings (54 in total) in MR while embodying two different avatar rendering styles. After each meeting, they answered a set of questionnaires.

In comparing the experiences of knowledge workers' using personalised realistic or cartoon avatars over multiple real-world meetings, we found that the avatar rendering style that they started with had an effect on their experiences, as did the time using the avatars. Overall, the study suggests that people have high expectations for the communicative and emotional value of realistic avatars, perhaps because they enable trust in the form of identification of the other, but that wanes quickly if avatars do not live up to expectations for other cues. A crucial finding was that avatars may be less effective for conveying negative emotions, especially realistic avatars. On the other hand, participants reported feeling more comfortable using cartoon avatars over time. A key message for future research and commercial usage, then, is to prioritise features and deployment plans around communicative value for the situations in which avatars will be used over accuracy (or perceived lack of accuracy) in likeness.

CRediT authorship contribution statement

Georgiana Cristina Dobre: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Marta Wilczkowiak:** Writing – review & editing, Visualization, Supervision, Conceptualization. **Marco Gillies:** Writing – review & editing, Writing – original draft, Supervision, Project administration. **Xueni Pan:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Formal analysis. **Sean Rintel:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Georgiana Cristina Dobre reports financial support was provided by Microsoft Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was conducted when the first author was employed at Microsoft Research Cambridge UK. We thank the Microsoft Mesh product group for their technical and research help. We also thank our participants for the generous contribution of their time and effort over the period of the study. Additionally, this work was partly supported by grant EP/L015846/1 for the Centre for Doctoral Training in Intelligent Games and Game Intelligence (http://www.iggi.org.uk/) from the UK Engineering and Physical Sciences Research Council (EPSRC).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2025.103632.

Data availability

The data that has been used is confidential.

References

Amadou, N., Haque, K.I., Yumak, Z., 2023. Effect of appearance and animation realism on the perception of emotionally expressive virtual humans. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. pp. 1–8.

Arboleda, S.A., Kunert, C., Hartbrich, J., Schneiderwind, C., Diao, C., Gerhardt, C., Surdu, T., Weidner, F., Broll, W., Werner, S., et al., 2024. Beyond looks: a study on agent movement and audiovisual spatial coherence in augmented reality. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 502–512.

Aseeri, S., Interrante, V., 2021. The influence of avatar representation on interpersonal communication in virtual social environments. IEEE Trans. Vis. Comput. Graphics 27 (5), 2608–2617.

Bailenson, J.N., Beall, A.C., 2006. Transformed social interaction: Exploring the digital plasticity of avatars. In: Schroeder, R., Axelsson, A.-S. (Eds.), Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments. Springer Netherlands, Dordrecht, pp. 1–16. http://dx.doi.org/10.1007/1-4020-3898-4_1.

Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M., 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence: Teleoperators Virtual Environ. 10 (6), 583–598.

Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M., 2003. Interpersonal distance in immersive virtual environments. Pers. Soc. Psychol. Bull. 29 (7), 819–833.

Bailenson, J.N., Yee, N., 2006. A longitudinal study of task performance, head movements, subjective report, simulator sickness, and transformed social interaction in collaborative virtual environments. Presence: Teleoperators Virtual Environ. 15 (6), 699–716.

Benford, S., Bowers, J., Fahlén, L.E., Greenhalgh, C., Snowdon, D., 1995. User embodiment in collaborative virtual environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 242–249.

Blomqvist, K., Cook, K.S., 2018. Swift trust: State-of-the-art and future research directions. In: The Routledge Companion to Trust. Routledge, pp. 29–49.

Brown, T., Smith, R., Zarate, D., Griffiths, M.D., Stavropoulos, V., 2024. Exploring user-avatar bond profiles: Longitudinal impacts on internet gaming disorder. Comput. Hum. Behav. 159, 108340.

Burgoon, J.K., Guerrero, L.K., Floyd, K., 2016. Nonverbal Communication. Routledge.

Collingwoode-Williams, T., O'Shea, Z., Gillies, M., Pan, X., 2021. The impact of self-representation and consistency in collaborative virtual environments. Front. Virtual Real. 2, 648601.

De Simone, F., Li, J., Debarba, H.G., El Ali, A., Gunkel, S.N., Cesar, P., 2019. Watching videos together in social virtual reality: An experimental study on user's QoE. In: 2019 IEEE Conference on Virtual Reality and 3d User Interfaces. VR, IEEE, pp. 890–891.

Dobre, G.C., Wilczkowiak, M., Gillies, M., Pan, X., Rintel, S., 2022. Nice is different than good: Longitudinal communicative effects of realistic and cartoon avatars in real mixed reality work meetings. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–7.

Félix-Brasdefer, J.C., 2015. The Language of Service Encounters. Cambridge University Press.

Fink, D.I., Skowronski, M., Zagermann, J., Reinschluessel, A.V., Reiterer, H., Feuchtner, T., 2024. There is more to avatars than visuals: Investigating combinations of visual and auditory user representations for remote collaboration in augmented reality. Proc. ACM Human-Computer Interact. 8 (ISS), 540–568.

Fraser, A.D., Branson, I., Hollett, R.C., Speelman, C.P., Rogers, S.L., 2024. Do realistic avatars make virtual reality better? Examining human-like avatars for VR social interactions. Comput. Hum. Behav.: Artif. Humans 2 (2), 100082.

Freiwald, J.P., Schenke, J., Lehmann-Willenbrock, N., Steinicke, F., 2021. Effects of avatar appearance and locomotion on co-presence in virtual reality collaborations. In: Mensch Und Computer 2021. pp. 393–401.

Fridman, L., 2022. Mark Zuckerberg: Meta, Facebook, Instagram, and the Metaverse. URL https://www.youtube.com/watch?v=5zOHSysMmH0. Feb 26, 2022.

Garcia, B., Chun, S., Kicklighter, C., Mai, B., Palma, M., Seo, J.H., 2021. Studying design attributes of virtual characters to support students' perceived experiences in virtual reality lectures.. Int. Assoc. Dev. Inf. Soc..

Hall, E.T., Birdwhistell, R.L., Bock, B., Bohannan, P., Diebold, Jr., A.R., Durbin, M., Edmonson, M.S., Fischer, J., Hymes, D., Kimball, S.T., et al., 1968. Proxemics [and comments and replies]. Curr. Anthr. 9 (2/3), 83–108.

Han, E., Miller, M.R., DeVeaux, C., Jun, H., Nowak, K.L., Hancock, J.T., Ram, N., Bailenson, J.N., 2023. People, places, and time: a large-scale, longitudinal study of transformed avatars and environmental context in group interaction in the metaverse. J. Computer-Mediated Commun. 28 (2), zmac031.

Han, E., Miller, M.R., Ram, N., Nowak, K.L., Bailenson, J.N., 2022. Understanding group behavior in virtual reality: A large-scale, longitudinal study in the metaverse. In: 72nd Annual International Communication Association Conference, Paris, France.

Harms, C., Biocca, F., 2004. Internal consistency and reliability of the networked minds measure of social presence.

Hartbrich, J., Weidner, F., Kunert, C., Raake, A., Broll, W., Arévalo Arboleda, S., 2023. Eye and face tracking in VR: Avatar embodiment and enfacement with realistic and cartoon avatars. In: Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia. pp. 270–278.

Heidicker, P., Langbehn, E., Steinicke, F., 2017. Influence of avatar appearance on presence in social VR. In: 2017 IEEE Symposium on 3D User Interfaces. 3DUI, IEEE, pp. 233–234.

Herrera, F., Oh, S.Y., Bailenson, J.N., 2018. Effect of behavioral realism on social interactions inside collaborative virtual environments. Presence 27 (2), 163–182.

Immohr, F., Rendle, G., Lammert, A., Neidhardt, A., Zur Heyde, V.M., Froehlich, B., Raake, A., 2024. Evaluating the effect of binaural auralization on audiovisual plausibility and communication behavior in virtual reality. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 849–858.

Immohr, F., Rendle, G., Neidhardt, A., Göring, S., Ramachandra Rao, R.R., Arevalo Arboleda, S., Froehlich, B., Raake, A., 2023. Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal VR communication. In: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences. pp. 209–215.

Jo, D., Kim, K.H., Kim, G.J., 2017. Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems. In: Proceedings the 30th Conference on Computer Animation and Social Agents. pp. 27–36.

Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E., 2022. Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision. Springer, pp. 345–362.

Khojasteh, N., Won, A.S., 2021. Working together on diverse tasks: A longitudinal study on individual workload, presence and emotional recognition in collaborative virtual environments. Front. Virtual Real. 2, 53.

Kim, H., Park, J., Lee, I.K., 2023. "To be or not to be me?": Exploration of self-similar effects of avatars on social virtual reality experiences. IEEE Trans. Vis. Comput. Graphics 29 (11), 4794–4804.

Koch, M., von Luck, K., Schwarzer, J., Draheim, S., 2018. The novelty effect in large display deployments–experiences and lessons-learned for evaluating prototypes. In: Proceedings of 16th European Conference on Computer-Supported Cooperative Work-Exploratory Papers. European Society for Socially Embedded Technologies (EUSSET).

Langa, S.F., Montagud, M., Cernigliaro, G., Rivera, D.R., 2022. Multiparty holomeetings: Toward a new era of low-cost volumetric holographic meetings in virtual reality. Ieee Access 10, 81856–81876.

Latifi, M.Q., Poulus, D., Richards, M., Yap, Y., Stavropoulos, V., 2024. Predicting proteus effect via the user avatar bond: a longitudinal study using machine learning. Behav. Inf. Technol. 1–17.

Latoschik, M.E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., Botsch, M., 2017. The effect of avatar realism in immersive social virtual realities. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology. pp. 1–10.

Lombard, M., Ditton, T.B., Weinstein, L., 2009. Measuring presence: the temple presence inventory. In: Proceedings of the 12th Annual International Workshop on Presence. pp. 1–15.

Lugrin, J.L., Latt, J., Latoschik, M.E., 2015. Anthropomorphism and illusion of virtual body ownership. In: ICAT-EGVE. pp. 1–8.

Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De la Torre, F., Sheikh, Y., 2021. Pixel codec avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 64–73.

MacDorman, K.F., Chattopadhyay, D., 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 146, 190–205.

Matthews, G., Jones, D.M., Chamberlain, A.G., 1990. Refining the measurement of mood: The UWIST mood adjective checklist. Br. J. Psychol. 81 (1), 17–42. http://dx.doi.org/10.1111/j.2044-8295.1990.tb02343.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1990.tb02343.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1990.tb02343.x.

Mayer, J.D., 1986. How mood influences cognition. Adv. Cogn. Sci. 290–314.

McDonnell, R., Breidt, M., Bülthoff, H.H., 2012. Render me real? Investigating the effect of render style on the perception of animated virtual humans. ACM Trans. Graph. 31 (4), 1–11.

Meyerson, D., Weick, K.E., Kramer, R.M., et al., 1996. Swift trust and temporary groups. Trust. Organ.: Front. Theory Res. 166, 195.

Moustafa, F., Steed, A., 2018. A longitudinal study of small group interaction in social virtual reality. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. pp. 1–10.

Nordin Forsberg, B., Kirchner, K., 2021. The perception of avatars in virtual reality during professional meetings. In: International Conference on Human-Computer Interaction. Springer, pp. 290–294.

Oh, C.S., Bailenson, J.N., Welch, G.F., 2018. A systematic review of social presence: Definition, antecedents, and implications. Front. Robot. AI 5, http://dx.doi.org/10.3389/frobt.2018.00114, URL https://www.frontiersin.org/article/10.3389/frobt.2018.00114.

Otto, O., Roberts, D., Wolff, R., 2006. A review on effective closely-coupled collaboration using immersive cve's. In: Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications. pp. 145–154.

Pakanen, M., Alavesa, P., van Berkel, N., Koskela, T., Ojala, T., 2022. "Nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in AR and VR based telexistence systems. Entertain. Comput. 40, 100457.

Pan, Y., Steed, A., 2017. The impact of self-avatars on trust and collaboration in shared virtual environments. PLoS One 12 (12), e0189078.

Parmar, D., 2017. Evaluating the Effects of Immersive Embodied Interaction on Cognition in Virtual Reality (Ph.D. thesis). Clemson University.

Paulhus, D.L., Bruce, M.N., 1992. The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. J. Pers. Soc. Psychol. 63 (5), 816.

Sakurai, S., Goto, T., Nojima, T., Hirota, K., 2021. Effect of the opponent's appearance on interpersonal cognition that affects user-to-user relationship in virtual whole-body interaction. J. Robot. Mechatronics 33 (5), 1029–1042.

Salagean, A., Crellin, E., Parsons, M., Cosker, D., Stanton Fraser, D., 2023. Meeting your virtual twin: Effects of photorealism and personalization on embodiment, self-identification and perception of self-avatars in virtual reality. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–16.

Shin, M., Kim, S.J., Biocca, F., 2019. The uncanny valley: No need for any further judgments when an avatar looks eerie. Comput. Hum. Behav. 94, 100–109.

Slater, M., 1999. Measuring presence: A response to the witmer and Singer presence questionnaire. Presence 8 (5), 560–565.

Smith, H.J., Neff, M., 2018. Communication behavior in embodied virtual reality. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Sonia, B., Suma, T., Agyemang, K., Oyekoya, O., 2023. Mapping and recognition of facial expressions on another person's look-alike avatars. In: SIGGRAPH Asia 2023 Technical Communications. pp. 1–4.

Standaert, W., Muylle, S., Basu, A., 2021. How shall we meet? Understanding the importance of meeting mode capabilities for different meeting objectives. Inf. Manag. 58 (1), 103393. http://dx.doi.org/10.1016/j.im.2020.103393, URL https://www.sciencedirect.com/science/article/pii/S0378720620303311.

Suma, T., Sonia, B., Agyemang Baffour, K., Oyekoya, O., 2023. The effects of avatar voice and facial expression intensity on emotional recognition and user perception. In: SIGGRAPH Asia 2023 Technical Communications. pp. 1–4.

Sun, Y., Won, A.S., 2021. Despite appearances: Comparing emotion recognition in abstract and humanoid avatars using nonverbal behavior in social virtual reality. Front. Virtual Real. 109.

Waltemate, T., Gall, D., Roth, D., Botsch, M., Latoschik, M.E., 2018. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. IEEE Trans. Vis. Comput. Graphics 24 (4), 1643–1652.

Weidner, F., Boettcher, G., Arboleda, S.A., Diao, C., Sinani, L., Kunert, C., Gerhardt, C., Broll, W., Raake, A., 2023. A systematic review on the visualization of avatars and agents in ar & vr displayed using head-mounted displays. IEEE Trans. Vis. Comput. Graphics 29 (5), 2596–2606.

Yoon, B., Kim, H.i., Lee, G.A., Billinghurst, M., Woo, W., 2019. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 547–556.

Yuan, L., Dennis, A., Riemer, K., 2019. Crossing the uncanny valley? Understanding affinity, trustworthiness, and preference for more realistic virtual humans in immersive environments. In: Proceedings of the 52nd Hawaii International Conference on System Sciences.

Zhang, Y., Yang, J., Liu, Z., Wang, R., Chen, G., Tong, X., Guo, B., 2022. VirtualCube: An immersive 3D video communication system. IEEE Trans. Vis. Comput. Graphics 28 (5), 2146–2156. http://dx.doi.org/10.1109/TVCG.2022.3150512.

Zibrek, K., Kokkinara, E., McDonnell, R., 2018. The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? IEEE Trans. Vis. Comput. Graphics 24 (4), 1681–1690.