

Social Interactions in Immersive Virtual Environments: People, Agents, and Avatars

Author: Georgiana Cristina DOBRE

Student ID: 33346662

Supervisors:

Prof. Xueni PAN

Prof. Marco GILLIES

Examiners:

Prof. Anthony STEED

Prof. Jonathan GRATCH

Goldsmiths University of London

Defended on 31st May 2023

Thesis submitted in requirements for the degree of

Doctor of Philosophy

Abstract

Immersive virtual environments (IVEs) have received increased popularity with applications in many fields. IVEs aim to approximate real environments, and to make users react similarly to how they would in everyday life. An important use case is the users-virtual characters (VCs) interaction. We interact with other people every day, hence we expect others to appropriately act and behave, verbally and non-verbally (i.e., pitch, proximity, gaze, turn-taking). These expectations also apply to interactions with VCs in IVEs, and this thesis tackles some of these aspects.

We present three projects that inform the area of social interactions with a VC in IVEs, focusing on non-verbal behaviours. In our first study on interactions between **people**, we collaborated with the Social Neuroscience group at the Institute of Cognitive Neuroscience from UCL on a dyad multi-modal interaction. This aims to understand the conversation dynamics, focusing on gaze and turn-taking. The results show that people have a higher frequency of gaze change (from averted to direct and vice versa) when they are being looked at compared to when they are not. When they are not being looked at, they are also directing their gaze to their partners more compared to when they are being looked at. Another contribution of this work is the automated method of annotating speech and gaze data.

Next, we consider **agents'** higher-level non-verbal behaviours, covering social attitudes. We present a pipeline to collect data and train a machine learning (ML) model that detects social attitudes in a user-VC interaction. Here we collaborated with two game studios: Dream Reality Interaction and Maze Theory. We present a case study for the ML pipeline on social engagement recognition for the *Peaky Blinders* narrative VR game from Maze Theory studio. We use a reinforcement learning algorithm with imitation learning rewards and a temporal memory element. The results show that the model trained with raw data does not generalise and performs worse (60% accuracy) than the one trained with socially meaningful data (83% accuracy).

In IVEs, people embody **avatars** and their appearance can impact social interactions. In collaboration with Microsoft Research, we report a longitudinal study in mixed-reality on avatar appearance in real-work meetings between co-workers comparing personalised full-body realistic and cartoon avatars. The results imply that when participants use realistic avatars first, they may have higher expectations and they perceive their colleagues' emotional states with less accuracy. Participants may also become more accustomed to cartoon avatars as time passes and the overall use of avatars may lead to less accurately perceiving negative emotions.

The work presented here contributes towards the field of detecting and generating nonverbal cues for VCs in IVEs. These are also important building blocks for creating autonomous agents for IVEs. Additionally, this work contributes to the games and work industry fields through an immersive ML pipeline for detecting social attitudes and through insights into using different avatar styles over time in real-world meetings.

Acknowledgements

This PhD journey has been an adventure and I could not have done it without the support of many people who kept me moving forward. First, I am deeply grateful to Sylvia Pan and Marco Gillies for being my supervisors for the whole PhD, and for their great insights, encouragement and support. I feel greatly privileged to have them guide me through this PhD journey. Thanks to them I got involved in many interdisciplinary projects and events, I met, worked and learnt from people of diverse expertise and I reached where I am today.

This thesis is very interdisciplinary and enriched by many collaborations. I would like to thank Sean Rintel and Marta Wilczkowiak from Microsoft Research Cambridge for their constant guidance and encouragements. Through the internship at MSR, I have started to understand the value of longitudinal studies, despite the challenging and laborious work involved. I extend my thanks to many others at MSR Cambridge from whom I learned and I got to exchange ideas. And finally, to all the participants who committed every day for a few weeks to work from inside a HoloLens2, despite the technical challenges and the peak holiday season.

Many thanks to Dave Ranyard for his insights and support even after finishing the internship with him. I couldn't have done one of the most technical and complex parts of this thesis without the people behind Dream Reality Interactive and Maze Theory. Their experience insights and open-mindedness to research practices were very valuable. Particularly, I'd like to thank Russ Hardling, Richard Bates, everyone from Dream Reality Interactive who welcomed me into their team (and their daily stand-up), and all the participants who volunteered their time to take part in my study.

I was fortunate to start my PhD in a collaboration with Antonia Hamilton's Social Neuroscience lab at UCL. I am very grateful for all the insightful meetings and suggestions, and for letting me work on a very rich dataset. I would like to extend my thanks to Jamie Ward and Patrick Falk, who helped enormously with the dataset and the project.

I am very grateful for the IGGI community who made the PhD journey less lonely, to those with whom I could share progress, concerns, research ideas and board-game nights. Particularly, many thanks to those from RHB328 during my first years, and the IGGIs in RCH during the last part of my PhD. I am also very appreciative of the financial support received from the IGGI CDT, the Rabin Ezra Fund and Google's Women Techmakers Scholarship.

I am very thankful for the support of my partner, Carlos. I was very lucky to have him by my side, especially during the difficult lockdown periods. I am grateful for the long conversations, for keeping me motivated and confident the many times when the impostor syndrome would kick in. Finally, I would like to thank my parents for their unconditional support and for being great listeners even when I was explaining my research in a mixture of Romanian and English.

Contents

List of Figures	6
List of Tables	7
1 Introduction	8
1.1 Motivation	8
1.2 Goals and contributions	9
1.3 List of publications	13
1.4 Research presentations and demos	14
1.5 Overview of next chapters	15
2 Literature Review	17
2.1 Understanding People - Social Interaction Theory	18
2.1.1 Monologue and dialogue	18
2.1.2 Grounding in communication	18
2.1.3 Gaze and turns	20
2.2 Building Agents	22
2.2.1 The sensing and responding loop	22
2.2.2 Virtual characters in different media	27
2.2.3 Social attitudes	30
2.3 Evaluating Avatars ' Appearance	32
2.3.1 Avatars in immersive virtual environments	32
2.3.2 Tasks and Environment Setting	35
2.3.3 Temporality in IVEs Communication	35
2.4 Industry Application and Ecological Validity	41
2.4.1 The game industry	41
2.4.2 Avatars in remote meetings	43
2.4.3 Ecological validity in IVEs	44
2.5 Summary of literature review	46
3 PEOPLE: Direct Gaze and the Frequency of Gaze Change	47
3.1 Introduction	47
3.2 Dyadic multimodal dataset	50

3.3	Data analysis and results	52
3.3.1	Gaze behaviour during conversational roles	52
3.3.2	The effect of being looked at on own gaze	53
3.4	Limitations and discussion	55
3.5	Summary	56
4	AGENTS: Immersive ML for Social Attitude Detection	57
4.1	Introduction	58
4.2	Challenges and Contributions	60
4.3	Method: social engagement detection	63
4.3.1	The scenario for data collection	64
4.3.2	User data collection in VR	66
4.3.3	Human annotations in VR	68
4.3.4	Annotations overview and validation	70
4.3.5	Questionnaire results	70
4.4	Training the detection component	71
4.4.1	ML algorithms	71
4.4.2	Proposed ML configurations	72
4.4.3	Input data for proposed ML training	73
4.4.4	Implementation	74
4.5	Results	75
4.5.1	Data post-processing	75
4.5.2	Model configurations	78
4.5.3	Derived vs raw features	79
4.6	Generalisation: ML pipeline for social attitude detection	82
4.7	Limitations and discussion	83
4.8	Conclusion	85
5	AVATARS: The Appearance Impact in Real-World Meetings	87
5.1	Introduction	88
5.2	Research Questions	89
5.3	Methodology	90
5.4	Data Analysis	94
5.4.1	Reverse Coding	94
5.4.2	Accuracy of perceived emotional states	94
5.4.3	Avatar Order	95
5.5	Results	95
5.5.1	Part One: Communication, Tasks and Presence	95
5.5.2	Part Two: Perceived and Self-Reported Emotional State	99
5.6	Discussion and Limitations	106
5.7	Conclusion	110
6	Conclusion & Future Work	112

6.1 Summarised contributions	112
6.2 Collaborative work	114
6.3 Limitations and future work	115
6.4 Outlook	117
6.5 Conclusion	118
A Questionnaires	135
B Study Information Sheets & Ethics Forms	146
C Monologue Script for Chapter 4	155
D Additional Stats for Chapter 5	159

List of Figures

1.1	Overview of the three main projects in this thesis	11
2.1	The sensing and responding loop	23
3.1	Setup for recording the dyadic data	51
3.2	Direct Gaze percent speaking and listening	53
3.3	Gaze dynamics while being or while not being looked at	54
4.1	Notes from brainstorming sessions with collaborators from the game industry.	61
4.2	Pipeline for detecting human-defined social engagement	63
4.3	Representation of the environment for data collection	64
4.4	Frontal view of the virtual character	65
4.5	Example of users interacting with the VC	66
4.6	Virtual Character’s view in both data collection batches	67
4.7	The expert’s annotation process	69
4.8	Accuracy of different ML models configurations using derived and raw data	80
4.9	Accuracy and F1 values for the PPO+GAL+LSTM+PreTrain model configuration trained with raw data	81
5.1	The Avatars and the Mixed Reality environment	89
5.2	Personalised full-body avatars	92
5.3	Responses to the questionnaire from Table 5.2 overall and over time	98
5.4	Self-reported emotional state ratings, overall, over time, by emotional state and by Order	100
5.5	Averaged error of the perceived emotional state by Order	102
5.6	Error of the perceived emotional state over time	103
5.7	Error of the perceived emotional state by self-rated emotional states	104
5.8	Ranking of the most useful emotional cues	106

List of Tables

2.1	Overview of previous work on avatars' styles	40
4.1	A snippet of VC's monologue.	67
4.2	Types of data recorded during the VC-user interaction	68
4.3	Socially meaningful derived data	73
4.4	The dataset used for the ML model training	74
4.5	Hyper-parameters and their values for the ML models	75
4.6	Confusion matrices for each ML model configuration.	79
4.7	Confusion matrices for PPO+GAIL+LSTM+PreTrain model configuration	82
5.1	Details on the participants and the data collected	92
5.2	The items in the daily questionnaire	95
D.1	P-values from the paired two-tailed t-test on the self-reported emotional state rating while embodying C or R avatars in order CR and RC (Cw1 vs Rw2; Rw1 vs Cw2)	159
D.2	P-values from regressions on the self-reported emotional states rating over time	159
D.3	P-values from regressions on the mapped error of the perceived emotional state over time	159

1

Introduction

1.1 Motivation

The use of immersive virtual environments (IVEs) has increased over the past years due to great technical advances making this technology more accessible and available on the commercial market. Consequently, these environments are more and more used in areas such as science, education, training, social and cultural experiences, travel, entertainment and news [Slater and Sanchez-Vives, 2016]. Of particular importance for developing virtual environments across many of these domains is the social interaction between humans and the characters within the virtual environments.

Virtual characters are a powerful tool for building these interactions in many applications for immersive virtual environments (Virtual Reality- VR, Mixed Reality- MR). Humans are engaging in social interactions every day developing strong observational skills which lead to being able to quickly notice errors in a virtual character's performance. These errors can be easily interpreted as the virtual character's attributes rather than technological flaws. Take for example incorrect gaze modelling: having a virtual character (VC) avoiding eye contact or not looking at the other person creates an impression of hostility and disinterest. Implementing a behaviour that makes it perform too much eye contact makes the virtual character appear too keen and unsettling. Moreover, poorly timed gaze behaviour, such as having the virtual character looking at the other person at the wrong time can disturb the conversation's smooth flow. Therefore, to create behaviour that transmits predetermined messages using an animated virtual character, developers have to carefully synchronise the virtual character's gestures with speech, facial expression and gaze with body posture, its actions with the contextual changes, and so on.

Research around low-level behaviour during human-to-human social interaction (such as modelling gaze shifts, head movements, gestures or speech turns) facilitate social interactions in IVEs. Hence, the development of virtual characters has applications in a range of areas such as therapy, training or entertainment. During social interactions, these low-level nonverbal behaviours contribute to implicit behaviour (that happens without conscious awareness) which has a role in driving social interaction. Another aspect lies in the interaction loop between the participants in an interaction. People adapt and adjust their verbal and nonverbal behaviours in day-to-day interactions based on their partner's behaviour and overall social interaction. Hence, during social interactions in virtual environments, the virtual character's non-verbal behaviour needs to take into account the user's and their own behaviour.

There are a few ways to generate and control a VC's behaviour in virtual environments. This can take place fully using algorithms (resulting in autonomous agents) or through a mixture of algorithms and human control (via Wizard-of-Oz, where a confederate partly drives the VC). When it comes to driving the VC using algorithms, there are a few types predominantly appearing in the literature. machine learning (ML) models are the ones that work in a diverse number of cases and topics, especially reinforcement learning and imitation learning as well as temporal models.

1.2 Goals and contributions

The work advances the area of developing virtual characters in IVEs from a point of view of detecting and generating nonverbal behaviours. A starting point for this is understanding interactions dynamic between *people*. This informs the development of virtual characters in IVEs. When it comes to virtual characters, they can be:

- *agents*, with no humans embodying them and being controlled mainly by computer scripts;
- *avatars*, with humans primarily controlling, embodying and being represented by them.

Hence, an *agent* is controlled by a computer script/algorithm[Bailenson and Blascovich, 2004]. If the algorithm controls everything about the agent (verbal and nonverbal behaviour), then they can be called 'autonomous agents'. However, this is a hard AI problem to date, as social interactions are very complex and differ from situation to situation. What happens instead is a middle ground, the agent is usually controlled partly by algorithms and partly by people. An example of this situation is the Wizard-of-Oz model, covered in many studies [Rizzo et al., 2015, Pan et al., 2012]. For instance, this is a Wizard-of-Oz system used Pan and Hamilton: an agent is performing basic, usually rule-based, nonverbal behaviours (i.e., gaze, blinks), and a human (usually a confederate) controls the higher-level nonverbal behaviours (i.e., hand/arm movements, body posture) and the verbal behaviour through a control panel with different buttons and sliders. When an agent is controlled by both humans and algorithms, they are usually called 'semi-autonomous agents' [Pan and Hamilton, 2018].

Usually, an agent is not represented as and identified by a certain person who is controlling it. For instance, in the work of Pan et al., the agent Christina is controlled by a confederate [Pan et al., 2012]. The confederate can be anyone who is trained to use the system, and they are not

identified with the character Christina. Whereas, in the work of Moustafa and Steed, the users embodied a virtual representation that was available on the social VR system. They could choose what representation to embody; over time they changed the representation to fit the interaction and the expectations of the rest of the group. [Moustafa and Steed, 2018] In this case, the user partly controls their virtual character- head movement, speech body and arms movement. Automatic scripts/algorithms control some other elements (blinks, eye movement, lip-flapping). Even though this seems very similar set-up to the semi-autonomous agents, these virtual characters are referred to as *avatars*. Ideally, an avatar will be fully controlled by the user embodying that avatar: the avatar will copy the human's exact behaviours (facial expression, body movement, gesticulations); and an agent will be fully controlled by computer algorithms (without the need for a human's input). However, this is not the case yet due to computational and technological limitations. Agents are partly controlled by humans (even though those who control them are not meant to embody them) and avatars are partly controlled by automated scripts.

Given this situation, the thesis's main focus is on virtual characters in IVEs. We focus on understanding the dynamics between *people* in social interactions to apply it to virtual characters in IVEs.

The initial aim of this work lies in advancing the area of autonomous agents in IVEs. As the autonomous agents are not triggered by humans, they need to *sense* (and consequently understand) what is happening in the interaction. After sensing, they need to *respond* to the information received. These two actions (sensing and responding) happen in a continuous loop, being the base of an autonomous agent. This loop is illustrated in Figure 2.1.

Building autonomous agents in immersive virtual environments can be very complex as there are many ways to approach this problem. Hence, creating fully autonomous agents is out of the scope because of the complexity of the problem and the time limitations of this PhD. Instead, we are contributing towards an autonomous agent and are focusing on social interactions in IVEs covering three interconnected areas: people, agents and avatars. Therefore, below are the **three main research questions** and the backbone of in this thesis:

- **Research Question 1:**

What are the dynamics of low-level non-verbal behaviours, in particular, gaze and turn-taking, in face-to-face social interactions between two **people**?

- **Research Question 2:**

How can an **agent** be trained to recognise implicit social attitudes during social interactions in virtual reality?

- **Research Question 3:**

What is the influence and response to the others' personalised **avatars** appearance during repeated social interactions between co-workers in mixed reality?

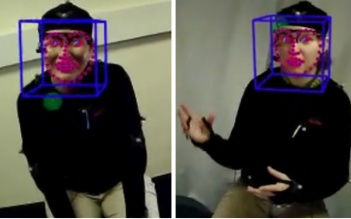


	People	Agents	Avatars
			
overview	interaction dynamics	social attitudes detection	appearance effect in meetings
contribution	automated speech and gaze annotation	immersive ML for data collection & annotation	cartoon vs realistic avatars in work meetings
ecological validity	unstructured tasks & multimodal dataset*	industry-focused project and applications	longitudinal & real meetings in own environment
medium & equipment	F2F w/ PupilLab glasses*	VR w/ Oculus Quest	MR w/ HoloLens2
collaborating bodies	Social Neuroscience Group, ICN, UCL	Dream Reality Interactive, Maze Theory	Microsoft Research Cambridge
outreach	ICMI2019	Springer VR, IVA2022	CHI2022

FIGURE 1.1: Overview of the three main projects in this thesis

Highlights of the three thesis chapters, showing the overview of the work, technical contributions, features of the ecological validity, the medium we run the study in, the equipment used, the collaboration bodies and the peer-reviewed venues where we published the results. *F2F* stands for *face to face*. *The dataset from the People study is a contribution from the Social Neuroscience group at the Institute of Cognitive Neuroscience from University College London.

RQ1. People’s conversational dynamics: gaze and turn-taking The low-level nonverbal behaviours that we focus on are gaze and turn-taking. These are important cues in social interactions. They are related to each other and they are core to social interactions. Gaze, for instance, manages the flow of the discussion, it is used to reference or to show interest in objects or other people, it can improve comprehension (in a teacher-learner situation), expresses complex emotions and facilitates interpersonal processes [Argyle and Ingham, 1972a, Mason et al., 2005, Bayliss et al., 2006].

There are shortcomings when implemented on two-dimensional (2D) displays, like on monitor screens. However, their roles are perceived more accurately once implemented in three-dimensional (3D) ones, such as IVEs. This is usually the case as gaze targets are more difficult to perceive on 2D screens compared to in 3D environments [Moubayed et al., 2012]. Gaze behaviours and turn-taking are intertwined in social interactions. Hence, the gaze targets perception affects turn-taking. For instance, in an interaction between a user and a VC, it is more challenging to convey the next

speaker on 2D displays compared to in 3D environments [Fischer and Tenbrink, 2003, Al Moubayed and Skantze, 2011].

Low-level non-verbal behaviours have a strong influence on social interactions hence an application that would take into account these signals can detect suitable non-verbal behaviours for the VC. By doing so, it can create an environment closer to the real-life one and hence, the user might have a higher sense of presence.

Although progress in the detection and generation of gaze and turn-taking are important contributions to the field, these cues are also conditioned by many other non-verbal cues or social aspects. Thus, they are difficult to generalise over, being influenced, for instance, by the relationship, user background interaction type or discussion topic. To design a study, these need to be controlled, hence they are not generalised for an autonomous VC.

In this work, we collaborated with the Social Neuroscience group at the Institute of Cognitive Neuroscience from University College London. They collected a multimodal dataset of dyads with free-flow conversations. We used this dataset in our study to understand the conversational dynamics in terms of gaze targets and turn-taking between two people.

RQ2. Agents recognising social attitudes Models for gaze and turn-taking help advance the field towards autonomous agents. Apart from these low-level cues, there are other nonverbal behaviours that are more abstract to define, for instance, social attitudes in an interaction. These higher-level nonverbal behaviours are more complex and they are impossible to describe with a straightforward set of rules. Examples of social attitudes include social engagement, sympathy, affection or aggression. These are more difficult to detect without rich multi-modal and socially relevant data. By allowing the VC to sense them (as humans do), the VC can respond appropriately to things that happen subconsciously/implicitly within the user.

This detection is useful in many application domains, especially in narrative VR games. In this situation, the VC is able to sense whether the user (or player) is showing a certain social attitude. Based on this, they could take actions that would fit that scenario. For this work, we collaborated with two game studios to build an ML pipeline that would detect a social attitude during a VR interaction. Our approach is based on ML models that have temporal features and that can learn from human examples.

RQ3. Avatars' appearance evaluation The last contribution tackles how the avatar's appearance influences the interaction in immersive virtual environments. The way a person is represented in virtual environments impacts the functional communicative value of the interaction, playing an important role in social encounters.

Setting aside the specific nonverbal behaviours, we focus on the overall avatar's representation. According to Benford et al., avatars represent people's identities, positions, interests, and activities [Benford et al., 1995]. Avatars can be represented in many ways, from floating spheres to full or partial humanoid bodies with various aesthetics (such as cartoon or realistic). Avatars may now be

extensively customised to closely resemble real people and adopt a particular look thanks to technological advancements. Different avatar styles have advantages and disadvantages. Using realistic avatars might cause people to feel uneasy and reduce their sense of affinity [Shin et al., 2019]. This is frequently caused by a mismatch between the avatar's behaviour and the high expectations of nonverbal behaviours (such body language and facial expressions). Generic or customised cartoon styling may make people question their appropriateness in a professional setting.

Additionally, a one-off interaction in virtual spaces might cause a novelty effect [Koch et al., 2018, Parmar, 2017]. This usually happens because users are not familiar with the new technology (such as headsets for IVEs). The task users do also plays a factor. In most cases, the task is distant from people's usual activities. Repeated exposure through a longitudinal study should tackle these aspects. However, they are not very common as they come with a high time and logistical cost.

For this study, we collaborated with Microsoft Research Cambridge. We investigated the effect of repeated usage of two avatar styles, photo-realistic and cartoon-like, during work meetings between two or three people in MR.

The results will inform how the repeated usage of different avatars will interact with the functional communicative value, task satisfaction, presence and the ability to identify other's emotional states.

In this thesis, we present contributions on these three elements: people, agents and avatars. First, we highlight insights into the social dynamics between **people**, then we cover how an **agent** could sense and recognise what is happening in social interactions with a user in VR in order to respond accordingly, and lastly, we evaluate the **avatars'** appearance in MR during repeated real-world meetings between co-workers.

1.3 List of publications

This thesis includes first-authored peer-reviewed material that has been or is under review to be published as follows:

Chapter 3:

Dobre, Georgiana Cristina, Marco Gillies, Patrick Falk, Jamie A. Ward, Antonia F. de C. Hamilton, and Xueni Pan. "Direct gaze triggers higher frequency of gaze change: An automatic analysis of dyads in unstructured conversation." *In Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 735-739. 2021 <https://doi.org/10.1145/3462244.3479962>

Chapter 4:

Dobre, Georgiana Cristina, Marco Gillies, and Xueni Pan. "Immersive machine learning for social attitude detection in virtual reality narrative games." *Virtual Reality* 26, no. 4 2022: 1519-1538 <https://doi.org/10.1007/s10055-022-00644-4>

Dobre, Georgiana Cristina, Marco Gillies, David C. Ranyard, Russell Harding, and Xueni Pan. "More than buttons on controllers: engaging social interactions in narrative VR games through social attitudes detection." *In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pp. 1-8. 2022 <https://doi.org/10.1145/3514197.3551496>

Chapter 5:

Dobre, Georgiana Cristina, Marta Wilczkowiak, Marco Gillies, Xueni Pan, and Sean Rintel. "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. 2022 <https://doi.org/10.1145/3491101.3519628>

1.4 Research presentations and demos

This section contains a list of the venues where we have presented formally the projects from this thesis. It contains talks, workshops, posters and demos. These cover largely initial results or proposed works. Although they are less rigorously reviewed than the published articles in conferences and journals, they have other advantages. First, they tend to be presented to the community in a more relaxed and practical manner, leaving room for discussions and further explorations of topics and ideas. They provide visibility to a wider audience, sometimes outside of the main study field. Furthermore, these allow raising awareness of our research and the gaps in the current work, inspiring further advancements in the field.

Talks

- "Almost humans in VR games: humanoids for social interactions", peer-reviewed, at the *Intelligent Games and Game Intelligence Conference*, London, UK, 2018
- "Virtual character's ability to recognise social behaviours in VR", at the *AI and Character Driven Immersive Experiences Workshop*, London, UK, 2019
- "AI-driven characters in VR, co-presented", at the *Immersive UK Webinar*, online, 2020
- "Implicit and engaging social interactions through social attitudes detection", co-presented, at *Develop Conference*, Brighton, UK, 2022

Posters

- "Non-Verbal Cues for Interactive Virtual Characters in Immersive Virtual Environments- Why Bother?", peer-reviewed, at the *Intelligent Games and Game Intelligence Conference*, London, UK, 2018

- *Mutual Gaze in Social Interactions*, peer-reviewed, at the *4th Workshop on Virtual Social Interaction (VSI)*, London, UK, 2018,
- "Non-verbal cues for interactive NPCs in VR games", peer-reviewed, at the *Intelligent Games and Game Intelligence Conference*, York, UK, 2019

Doctoral Consortium - each including a talk and poster presentation

- "Using machine learning to generate engaging behaviours in immersive virtual environments", peer-reviewed, at the *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, UK, 2019
- "People, Agents, and Avatars: nonverbal behaviours in immersive social interactions", peer-reviewed, at the *22nd Conference on Intelligent Virtual Agents (IVA)*, Faro, Portugal, 2022

Workshops

- "Interactive ML for VR Game Controls", co-run, peer-reviewed, at the *Intelligent Games and Game Intelligence Conference*, London, UK, 2018
- "Imitation Learning for Unity Games", co-run, peer-reviewed, at the *Intelligent Games and Game Intelligence Conference*, York, UK, 2019
- "Immersive Interaction design for IVAs", co-run, peer-reviewed, at the *20th Conference on Intelligent Virtual Agents (IVA)*, online, 2020

Demos

- "Virtual agents recognising social engagement in VR", at the *AI and Character Driven Immersive Experiences Workshop*, London, UK, 2019

1.5 Overview of next chapters

The following Chapter 2 will cover the social interactions theory with a focus on communication, gaze and speech turns. It will include insights into the nonverbal behaviour dynamics in interactions and different models of gaze behaviour and speech turns. It will include the background on building agents, evaluating avatars and the implications of these in different industries, along with aspects of ecological validity.

Next, in Chapter 3 we will introduce the first study on the nonverbal behaviour dynamics in dyadic social interactions, focusing on conversational turns and gaze. This is based on a dataset on dyadic interaction during an unstructured and free-flow conversational task. The results bring insights into

the conversational dynamics of speech turns and gaze targets that will contribute to building agents for IVEs.

Then, in Chapter 4 we present an Imitation Learning approach to developing an ML framework to detect complex attitudes in social interactions in VR. We give the case study of social engagement in a VR environment build in partnership with and for two game studios.

In Chapter 5 we introduce a longitudinal study on the effect of avatars' appearance in mixed reality work meetings. Here, we cover the study design, research questions and the data collected to compare the effect of avatars in cartoon and realistic styles.

This will be followed by Chapter 6 where we will conclude with the future work on the contributions of this thesis, the implications and the impact of the work proposed.

2

Literature Review

The area of social interactions is complex, and to get a better understanding of the underlying mechanics of social interaction is essential to look into the research and discoveries from this and related disciplines.

This chapter starts with understanding **People** by introduction: grounding in communication, theory and the state of the art on gaze and turn-taking (Section 2.1).

In the next part, we cover some of the elements of building **Agents**: how is the data collected and used to generate behaviour for agents (with a focus on gaze and speech turns); what is the state of the art on virtual characters for 2D displays; and concluding with the most up to date agents for IVEs and the gaps in the field (Section 2.2).

After this, we focus on a different area regarding social interactions: the evaluation of the virtual representations or appearance of **Avatars** in IVEs (Section 2.3). Here we bring insights into the avatar's body, style, resemblance to the user and avatar usage (one-off interaction or repeated, long-term one).

Next, we include how are the virtual characters and avatars used in the industry covering two main industries: gaming and remote working (Section 2.4). This brings into the discussion the ecological validity of the state-of-the-art existing work and their applicability in the real world. We conclude this chapter by summarising the literature review (Section 2.5).

2.1 Understanding People - Social Interaction Theory

2.1.1 Monologue and dialogue

A dialogue can pose a number of difficulties that have to be addressed for an effective conversation. For instance, it could be difficult to predict how the conversation will unfold and therefore, people are not able to plan what to say far in advance; or decide who to address in a multi-party conversation. Other problems are knowing when it is socially appropriate to speak, planning what to say while still listening to the partner, dealing with incomplete and very brief conversational utterances and so on. Despite all of these difficulties, a dialogue is still easier than a monologue because, as Clark describes, it is a joint process [Clark, 1996]. The conversational partners work together to form a mutual understanding of what they are discussing. They are also aligning their representation of time, space, causality and intentionality, in short, their representations of situation models [Garrod and Pickering, 2004]. Garrod and Pickering argue that this process, which they call ‘interactive alignment’, takes place unconsciously, automatically, and it reinforces the fact that humans are designed for dialogues rather than monologues [Pickering and Garrod, 2004].

2.1.2 Grounding in communication

When people are taking part in a conversation they coordinate both the content of the conversation and the process. They start coordinating the content before starting the conversation by estimating their *common ground*: mutual knowledge, beliefs and assumptions [Clark and Marshall, 1981, Clark and Carlson, 1982, Clark and Schaefer, 1987, Clark and Schaefer, 1989]. They tend to coordinate their communication process by referring back to the updates on their common ground [Clark and Brennan, 1991]. Grounding is a process introduced by Clark and Schaefer, to illustrate how, during a conversation, the common ground is changing from moment to moment [Clark and Schaefer, 1989]. During a conversation, people are trying to determine if what they said was understood, or if what they said is now part of the participants’ common ground. To do this, people are looking for evidence in grounding, evidence that shows the speaker was heard and understood. This can be negative (expressing an utterance misunderstanding) or positive (expressing comprehension of what was said). From these two types, the positive evidence is the most sought one, the most common forms employed by the addressees are acknowledgements (or back-channels), relevant next turn or continued attention.

Evidence in grounding. People show their understanding of an utterance using different methods, such as acknowledgements, continuing with a relevant next turn, or expressing a continuing attention towards the speaker. Acknowledgements include back-channels responses (such as *uh huh, yeah*), assessments (like: *really, oh God*) [Goodwin, 1986] and various gestures (such as head nods) [Goodwin, 1981]. An example of a relevant next turn is the answer to a question which forms a so-called adjacency pair. After the first part of this adjacency pair is finished (an asked question), the second part is a conditionally relevant next turn (the answer). In this case, the turns are not only utterances but answers that are directly related to the question, hence the evidence of understanding/of grounding is much stronger [Schegloff and Sacks, 1973]. Finally, the last way to express

evidence in grounding is by continuing the attention stream to the speaker. This is the most basic form of showing understanding when communicating. People tend to closely monitor their partners and also observe what they are attending to [Goodwin, 1981]. In this way, the listener accepts the information given in the conversation.

The least collaborative effort. In grounding, people try to minimise effort from themselves and from the conversational partner. This principle is known as the least collaborative effect [Clark and Wilkes-Gibbs, 1986]. This effect is further described as the participants in the conversation trying to minimise their work starting with the initiation of each utterance to mutual acceptance of an ended utterance.

Grounding and communication purpose. In conversations, people automatically set a common purpose, common topic of discussion, and a certain type of content is expected to be discussed [Grice, 1975]. Special techniques have evolved for grounding certain pieces of conversation. It is important in a conversation to have a reference well established and people use techniques that are designed for that purpose. For instance, many conversations focus on objects and their identities. People often need to identify and reference those objects quickly and effectively.

There are a few ways people reference objects [Clark and Brennan, 1991]. It can be through providing an alternative description of a new object using information that is already in the common ground. Or by using indicative gestures to point to the new object that needs to be identified. Another way is by using a trial reference: speakers might present a description of a new item/object and if they are not entirely sure about its comprehensiveness, they might ask their conversational partner to confirm their understanding of the new information before finishing/continuing the utterance [Clark and Brennan, 1991].

There are different ways of grounding in the case of referencing verbatim content. The most commonly used ones are: repeating the verbatim, spelling the information or, in the case of an address or long number, breaking the information into manageable chunks (instalments). The former is often carried out as people have limited immediate memory spans. In a study on calls to directory inquiries by Clark and Schaefer, the operators always divide the phone numbers into conventional groups [Clark and Schaefer, 1987].

Grounding, communication mediums and its constraints. As there are techniques for grounding different communication purposes (as seen above), grounding can be also different based on the conversation medium. People tend to ground with those techniques available in a medium that would eventually lead to the least collaborative effort [Clark and Wilkes-Gibbs, 1986]. There are a number of constraints on grounding that vary for each medium. A few examples of constraints are: co-presence (having the participants share the same physical space), audibility (having the participants communicate by speaking) or reviewability (being able to review a message). Face-to-face medium benefits co-presence and audibility but, in this medium, the messages are not reviewable.

For instance, answering machines benefit from audibility and reviewability but lack co-presence whereas letters and electronic mails are reviewable but they don't benefit from audibility or co-presence.

Grounding costs. As mediums are differently constrained, people are forced to use alternative grounding techniques, these being further balanced based on grounding costs. A *grounding cost* refers to the cost the participants would pay in communication to ground their conversation. A few examples of these costs are: *formulating costs* (time and effort to plan and formulate an utterance), *start-up costs* (cost for starting a conversation, of getting the conversation partner to notice and accept your utterance), *delay costs* (cost of delaying an utterance in order to plan it; this could be also mistaken for a finished turn if the pause takes place in the middle of an utterance); other are speaker change costs, fault costs or repair costs.

In conclusion, in a conversation, grounding changes with both the content and medium. People set the common ground and communicate effectively in most mediums by taking into account the techniques available in the medium, the constraints and by carrying out cost trade-offs. As seen above, grounding is a core element in social interactions between people, and the same concepts apply to social interactions in immersive virtual environments. Making use of the relevant grounding elements when building a virtual character helps create impactful social interactions in these virtual environments. For example, interactions between users and agents are smoother if there is appropriate evidence in grounding. This can be done by providing suitable back channels (*yeah, uh huh*), head nods, assessments (*oh, really*) or even appropriate eye contact when listening and averting gaze when processing new information and attempting to take the next turn. If grounding is missing, the interaction flow gets disrupted. For instance, this can happen when in a conversation with a user the agent gazes or gesticulates toward an area or object that is not relevant to the discussion. This situation is likely to happen if the agent has pre-recorded nonverbal behaviour (body movement, gaze) that is not contextual and adaptable to a large range of interactions. As gaze and turn-taking are important nonverbal cues for establishing grounding, next we cover the theory and early work on gaze and turn, and how these social signals are essential building blocks for virtual characters in IVEs.

2.1.3 Gaze and turns

Gaze directions and utterance occurrences. Gaze direction has a great social significance, Tompkins reviewed the early writers on this topic [Tompkins, 1963]. The direction of gaze can have different functions in social settings. Kendon performed an exploratory study on the relationship between gaze direction and turn-taking in the context of an ongoing conversation between two people with a focus on two functions of gaze direction [Kendon, 1967]. First is the function of perception, by which a person in a conversation can monitor the other person's behaviour. Second, the gaze direction is seen as a sign of expressiveness and regulation in a conversation, by which one influences the behaviour of the other people in the conversation.

Changes in conversational turns usually take place when the speaker passes their turn to their partner; these are signalled by low-level and easy-to-observe behaviours, such as vocal and lexical features, head and eye direction [Duncan, 1972]. Direct and averted gaze behaviours are also linked with speech and turn transitions and are used as signals for different social intentions [Sandgren et al., 2012]. Speakers tend to start their turn with averted gaze and to end it with direct gaze, possibly to signal the turn is ending and to yield it to the other person [Kendon, 1967, Cummins, 2012, Duncan and Fiske, 1977].

Frequency and duration of averted and directed gaze. During conversations, the amount of time one person looks at the other one varies considerably. In Kendon's work it ranges from 28% to 70% [Kendon, 1967]. Similar discrepancies are also found in Exline's or in Nielsen's work [Exline, 1963, Nielsen, 1964]. Kendon notes that, on average, in a conversation during listening, people look at the speaker for longer, these direct gazes being broken only by very short averted gaze periods [Kendon, 1967]. However, the direct and averted gaze behaviour of the person who has the turn (person who speaks) tends to be, on average, more equal in length. In this case, the averted gaze is considerably longer than the listener's averted gaze.

Kendon also notes that the mean period of time the speaker looks at the listener and its gaze direction change rate are closely related to the other person's direct gaze and the corresponding change in gaze direction [Kendon, 1967]. In each dyad from this study, the participants coordinate their gaze behaviour, hence the proportion of direct and averted gaze changes from dyad to dyad. Kendon also suggests that the behaviour of one looking at the other is influenced by the dyad's participants. In one example where participants were discussing in dyads, one participant (P1) took part in a conversation with two other different participants (P2, P3). In these conversations, the mean direct gaze of P1 toward P2 and of P2 toward P1 was half of the mean directed gaze of P1 towards P3 and P3 towards P1.

Gaze behaviour considering long and short utterances. A long utterance or a speaking turn usually requires one person to speak while the other remains silent for almost the whole duration of the turn. Often, these involve the participant planning their speech in advance of starting the long utterance. The short utterances, however, are prompt reactions to the other person's ongoing behaviour, requiring no planning phase. They are speech habits (such as back-channels, and assessments), exclamations, attempt interruptions or very short questions. In many cases, these are produced during the other person's long utterance.

Kendon suggests that there is a gaze direction pattern at the beginning and at the end of long utterances [Kendon, 1967]. Results show that 76% of gaze is averted from the listener as the speaker starts their turn or in advance of starting their turn. And during the last sentence, as the turn ends, 78% of the gaze is directed to their interlocutor. This is also suggested by the fact that the person who is about to start a turn would not be able to effectively plan their speech and monitor

the interlocutor's behaviour at the same time.

The direction of gaze during long utterances is also influenced by the speed of speech or by whether a speech section is fluent or hesitant. During a hesitation, the speaker might be planning the next speech section and averting the gaze would allow them to do so without having to simultaneously audit the other person's behaviour. In Kendon's study, 72% of speech rate was faster during direct gaze than during averted gaze, and 26% of speech rate was slower while direct than during averted gaze [Kendon, 1967]. The proportion of fluent speech during direct gaze was also higher - 50%, whereas the proportion of hesitant speech was only 20.3%.

During long utterances, the listener might interrupt the speaker by performing a short utterance. Such short utterances influence the gaze direction [Kendon, 1967]. For example, during short questions (these are usually asked when one needs more information on the current topic) the person who asks the question looks straight at the other person while asking it; if the answer is short and prompt, the person who answers continues looking at the other person while answering, whereas if they are hesitant about the answer, it is likely their gaze will move away from the interlocutor.

Gaze direction can also signal the preferred response to a question. Polar questions are structured in a distinct way to imply a specific response, either yes or no. In their study, Kendrick and Holler. analyse the relationship between gaze direction and the preferred response. The results show that 82% of dis-preferred responses are answered with averted gaze and 60.2% of preferred responses are answered with directed gaze [Kendrick and Holler, 2017].

Another factor in gaze is an approach/avoidance conflict. Argyle and Dean propose that eye contact or mutual gaze (when both participants are gazing at the other) is, on one hand, actively sought in conversations for increasing closeness and self-validation [Argyle and Dean, 1965]. On the other hand, there is a tendency to avoid excessive mutual gaze, as it can be overly intimating and arousing. This leads to a conflict that is normally resolved by reaching an equilibrium level of mutual gaze.

2.2 Building Agents

2.2.1 The sensing and responding loop

Based on the literature on early work (see Chapter 2.1), the social interactions between people happen taking into account a loop of two actions: *sensing* and *responding*. For example, in a dyad, one person senses what is happening in the interaction and acts (responds) accordingly. The other person perceives the action (senses the change in the interaction) and responds based on that too. The first person senses (again) the response and takes action as expected. The loop continues this way until the end of the interaction. Hence, in order to build an agent that will be able to take part in social interactions with users, they need to *sens* what is the current state of the interaction

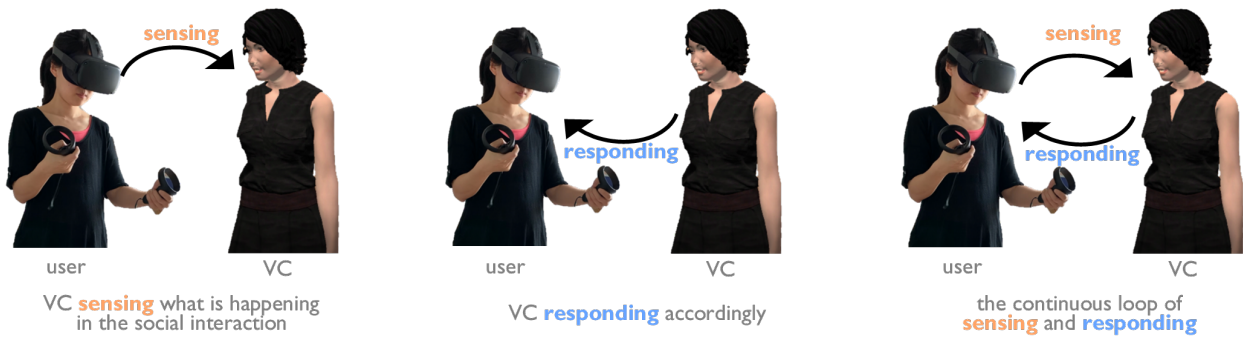


FIGURE 2.1: The sensing and responding loop

Representation of the sensing and responding loop in social interactions between a dyad. First, the virtual character (VC) **senses** what is happening in the interaction with the user. Based on this information that the VC senses from the user and the environment, the VC **responds** as expected by the user. This loop continues throughout the social interaction.

and then to *respond* as expected by the user. The sensing and responding loop is also illustrated in Figure 2.1.

To develop the *sensing* part of the loop for an agent, first, we collect data from the user, we annotate it and make sense of it. Based on this first process, the agent responds to the other person by a generated behaviour (verbal and nonverbal). Further in this section, we cover the data annotation practices (including manual and automatic annotations) which are the building blocks for the *sensing* part in the loop. Then, we detail the *responding* part, introducing different methods of generating nonverbal behaviour, what are their advantages and shortcomings, and when it is best to use them. Finally, we cover the state-of-the-art models for generating conversational cues.

Data annotation

To gather insights into social interactions and nonverbal behaviours, different kind of data is collected and analysed. Nonverbal and verbal behaviour data can be collected using different types of technologies. Usually, this data is recorded in its *raw* format. For instance, gaze data can be recorded using different types of eye-tracking devices. The outcome from the recording of eye behaviour using this technology is usually different video streams from the user’s view, and some dots (via pixels locations) representing where the user is looking at. This data can be also exported as pixel coordinates of the gaze. Similarly, speech data is recorded as audio files and their acoustic features (i.e., amplitude, pitch) represent the raw data.

This kind of raw data needs to be analysed in order to extract information from it. There are many ways of annotating data. The most common one is to manually label it. Usually, a few people go through the same data individually. They manually mark when a certain action happens such as when a participant speaks, when two participants look at each other, and so on. After each person annotates the data, the annotations are usually compared for accuracy, a widely used method is calculating the Cohen’s Kappa [Cohen, 1960]. Cohen’s Kappa is an agreement index that takes into account the agreement occurring by chance.

Manual annotations are time-consuming and challenging to scale to large datasets. Although there are no automatic solutions for all nonverbal behaviours, there are some alternatives for speech for instance. The field of speaker segmentation looks at ways of determining where there is a change in speakers from audio data [Kotti et al., 2008]. This is not an easy task, as the audio could contain backchannels or overlapping speakers, known as the cocktail party effect [Haykin and Chen, 2005]. The methods of speaker segmentation usually fall into two categories: metric-based or classification-based. The metric-based approach works by using a sliding window on the audio data and measuring the similarity between the adjacent sub-windows of the data within it at each window positioning. If the level of similarity is below a certain threshold, then it registers a speaker change [Dhananjaya and Yegnanarayana, 2008, Tritschler and Gopinath, 1999, Yang et al., 2005]. The classification-based approach is usually a binary classifier (i.e., support vector machine, neural networks) trained to detect talking [Gupta, 2015, Wang et al., 2017].

There is some prior work automatically annotating gaze targets in social interaction using video and/or gaze-tracking information. For instance, McLaren et al., describe a heuristic-based approach to detect gaze in a three-party dialogue [McLaren et al., 2020] using video data and the OpenCV library [Bradski, 2000]. Another example could be from the computer vision field, where Recasens et al. predict where someone is looking in videos, including the cases where the gaze target is not in the same frame [Recasens et al., 2017].

Generating nonverbal cues

Literature suggests that in most cases, VCs' nonverbal behaviour is generated through statistical modelling, rule-based or by making use of machine learning (ML) models. These approaches lead to semi-autonomous (partly controlled by a human) or autonomous VCs.

Statistical modelling refers to generating nonverbal behaviour based on conditional probabilities observed in human-to-human interactions, based on results from existing literature [Kendon, 1967, Argyle and Ingham, 1972a]. For example, the VCs can have set percentages for looking and not looking at the user, categorised by the speaking/listening mode they are in [Lee et al., 2002].

Rule-based methods use simple hard-coded algorithms, the VC performing a behaviour based on predefined if-then rules [Marsella et al., 2013]. Nonverbal behaviour is often pre-captured (body and face motion capture, eye tracking) and then played back based on these rules. These are more expressive, however, the behaviour is limited to what was captured offline, not taking into account the user's real-time feedback in the interaction. Pre-recorded motion data can be incorporated with a Wizard-of-Oz system, where an assistant operates or partially operates the VC [Pan and Hamilton, 2018, Pauw et al., 2022]. In most cases, a mixture of these methods is used to represent the VC's nonverbal behaviour.

ML models use pre-recorded data to learn nonverbal behaviour which is then applied to the VC. A large number of ML models show good results in predicting and generating nonverbal behaviour [Morency et al., 2008, Chiu et al., 2015, Haag and Shimodaira, 2016, Ferstl and McDonnell, 2018, Greenwood et al., 2017]. For instance, in Ferstl and McDonnell's work, features from an utterance

are fed into a recurrent network with an encoder-decoder structure to generate a sequence of gesture motions [Ferstl and McDonnell, 2018].

Nonverbal behaviour based on tasks. The methods described above can perform differently based on the nature of the interaction. Usually, this is determined by the task participants do, which can be structured and unstructured. In an interaction with *structured tasks*, the user performs set actions (such as changing the location of a cube based on a given instruction [Nguyen et al., 2018]). Interactions with *unstructured tasks* are more flexible. They don't have a pre-defined structure or some fixed actions to be completed. An example of an unstructured task is the free-flow conversation. Although there might be a pre-defined topic of conversation, people in interactions with unstructured tasks are not given precise instructions on when to be a speaker or a listener.

Statistical modelling works well for generating specific nonverbal cues (looking away/looking at the partner) in structured tasks.

Unlike statistical modelling which looks at specific low-level behaviours, rule-based models often use pre-recorded animation and can provide expressive whole-body nonverbal behaviour that fits into a structured task. This results in a more coordinated behaviour for the VC, but limits it to what is pre-recorded, making it less suitable for unstructured tasks. Both methods lack flexibility (able to adapt to different changes) and are not contextual (able to relate to complex circumstances).

ML models address some of these drawbacks. When trained on unstructured tasks, the ML models can be applied to a variety of scenarios; whereas ML models trained on structured tasks are limited to the types of scenarios they could be applied to. An example is the work of Nguyen et al, where in the structured task, the user moves a cube from one location to another based on the instructions received [Nguyen et al., 2018]. The complex ML model that generates head motions is built on head orientation along with task-related features (such as the user's hand location: grasp/move cube). Although the results are good, the model's features are task-related, making it unable to be re-applied to other scenarios.

The interaction between the VC and the user is not predictable. Hence, statistics and rule-based models do not perform well on unstructured tasks as not all possible events can be covered and foreseen. In an unstructured scenario, the VC-user rapport plays an important role. An example is in the IVE applications used for general practitioner training [Pan et al., 2016, Pan et al., 2018]. Here, the nonverbal behaviour is strongly linked to the empathic social interaction, the doctor's gaze and body orientation influencing the viewer's perception of their empathy [Brugel et al., 2015]. Another example is the use of virtual humans to provide socio-emotional support. Pauw et al. found that participants felt an improvement after talking to a virtual human, in terms of emotional and cognitive support [Pauw et al., 2022].

Various ML models have been used to generate nonverbal behaviours [Chiu et al., 2015, Haag and Shimodaira, 2016, Ferstl and McDonnell, 2018]. However, these are based on the data captured from a single person (for instance, gestures are generated from only their own speech without considering the interaction with another person [Ferstl and McDonnell, 2018]). This is an important aspect as a person's behaviour is different in the presence of another person, compared to when they are left

alone [Schilbach et al., 2013]. Thus, using ML models trained on self-data to generate behaviour for VC might result in non-engaging, less dynamic interaction, that lacks empathy. To tackle this, ML models could be based on the synchronised data from all participants in the interaction. The interaction dynamics could be captured, and the model could be able to generate engaging nonverbal behaviour, driven by the interaction between the participants. This approach is successfully used by Hoegen et al. to update an agent’s conversational style based on the one the interlocutor is using [Hoegen et al., 2019].

The methods presented perform differently based on the type of interaction they are applied to. In scenarios with structured tasks (where the user’s actions are limited to specific ones), they tend to perform well. However, in free-flow scenarios, with no pre-defined structure or no fixed actions, statistical modelling and rule-based models struggle. For these cases, ML methods tend to show good results [Forbes-Riley et al., 2012, Dermouche and Pelachaud, 2019a].

Models for conversational cues. So far we covered the background theory and how to analyse the collected social interaction data. The next step is using these results to build nonverbal behavioural models for an autonomous agent. With these models, the autonomous agent can automatically perform context-relevant non-verbal behaviours. Although this seems straightforward, it is still an open challenge to develop animated models that can map between the agent’s context-relevant emotional states and its gaze or speech behaviour (or other nonverbal behaviours). There are a number of approaches proposed in the literature and a selection of them is described next in this section.

Park Lee et al. define an eye movement model that is based on empirical models of saccades and statistical models of eye-tracking data [Lee et al., 2002]. The model is built on three major components. First, the Attention Monitor component monitors the system’s current state (such as changes in listening/speaking modes, gaze and head orientation and so on) and based on this, invokes one of the other two components: Parameter Generator or Saccade Synthesizer. The parameter Generator component decides on the duration of direct or averted gaze and on saccade magnitude, velocity or duration. Finally, the Saccade Synthesizer component collects the synthesizer parameters from the previous component and generates the sequence of coordinates value for the eye movement.

Andrist et al. describe a hybrid stochastic and heuristic bidirectional gaze mechanism which is synthesized from data collected from human-to-human interactions [Andrist et al., 2017]. It is built on top of an additional gaze model that is able to execute gaze shifts from one target to another [Andrist et al., 2012]. The bidirectional model has two main components. First, a stochastic part containing statistical parameters about what object/direction to gaze at and when to perform this gaze action. And second, a heuristic rule-based component that defines what the agent should gaze at in response to the information gathered from monitoring the user’s actions.

Zoric et al. present a method for automatically adjusting facial gesturing for virtual agents based

on their speech [Zoric et al., 2011]. Facial gestures include nods, head movements, blinks, eyebrow movements, and gaze. The method is based on a hybrid data-driven and rule-based approach and it works in three consecutive steps. First, it uses a Hidden Markov Model to classify the facial gestures into four main categories (blink, eyebrow movement, nod and swing). Next, the results are further defined by rules generated from information on non-verbal communication (from the literature). Finally, the facial gesture is fine-tuned assigning a gesture sub-type, amplitude and duration based on the statistics gathered from a training dataset.

Bohus and Horvitz introduce another approach to create a gaze model from speech [Bohus and Horvitz, 2010]. In their study, they design gaze shifts for agents to shape a multi-party conversation. The agent shows their turn-taking intentions by directing their gaze towards specific participants and averting it from other ones.

Le et al. also use speech inputs to generate gaze, eyelid motion and head motion [Le et al., 2012]. In their study, statistical models for each component are learnt separately. Such as, the head motion is generated using Gaussian Mixture Models and gradient descent optimisation algorithm. The non-linear Dynamic Canonical Correlation Analysis model is used to create gaze behaviour from both head motion and speech features. Finally, for generating eyelid motion, they employ non-negative linear regression.

A rule-based system is proposed by Marsella et al. that determines the semantic, pragmatic and rhetorical content of the utterance by using a shallow analysis [Marsella et al., 2013]. This analysis is then used to generate head movements, eye saccades, gestures, blinks and gazes.

Vinayagamoorthy et al. propose a method for designing a gaze model for immersive virtual environments [Vinayagamoorthy et al., 2004]. This model is designed for user-controlled virtual characters involved in dyadic verbal communications. The mechanism is built based on data and using prior results from studies on the behaviour of eye gaze while face-to-face social interactions.

2.2.2 Virtual characters in different media

There are a large number of autonomous and semi-autonomous virtual characters that are deployed on non-immersive 2D screens (monitors). Some of these are also available for immersive virtual environments, with some limitations. Below is a description of the more developed ones. They could act as the base for further development and further integration into immersive virtual environments.

Based on the SEMAINE API [Schröder, 2010], the Sensitive Artificial Listeners (SALs) are virtual characters with very limited verbal understanding [Kenny et al., 2007]. There are four SALs, each with a different personality: aggressiveness (Spike), cheerfulness (Poppy), gloominess (Obadiah) and pragmatic (Prudence). There are four SALs prototypes developed. The first version is the

PowerPoint SAL. For this, an operator chooses a sentence from 20-30 options and delivers it in an appropriate tone of voice. The Semiautomatic SAL uses the same concept as in the first version with the improvement of using pre-recording audio of the chosen sentence. Solid SAL is similar to PowerPoint SAL with the difference of fewer options to pick from when selecting a response. This tries to shift the focus on the operator's speech to be as relevant as possible to the character's personality. The main aim of this prototype is to train emotion recognition and feature detection systems. Finally, the Autonomous SAL is built on the previous versions. Even with limited domain knowledge, it is able to engage in interactions with a participant without being directly controlled by a human. The system analyses the user behaviour, manages the dialogue with the user, and synthesises its speaking and listening behaviours. The speaking behaviour mostly covers utterances that prompt the user to say more or are determined by the SAL's "goal" to draw the user toward the character's emotional state [Kenny et al., 2007].

Researchers from ArticulateLab at Carnegie Mellon University developed a number of virtual characters studying human interaction in social and cultural contexts. Of particular interest is the Socially-Aware Robot Assistant (SARA) [Matsuyama et al., 2016]. SARA is a virtual character that achieves a task or a social goal by expressing visual, vocal and verbal behaviours using multimodal user data. Data includes the user's face and head movements, vocal features, and conversational strategy. SARA is designed to raise or maintain the level of rapport with the user over the course of a conversation by understanding the information extracted and generating visual, vocal, and verbal behaviours. SARA's architectural components are built on top of the Virtual Human Toolkit [Hartholt et al., 2013], additional modules include Microsoft's Cognitive Services API for converting speech to text, Microsoft's LUIS (Language Understanding Intelligent Service) for recognising the user intent, OpenFace and OpenSmile for extracting facial features, head pose, gaze and acoustic features from the audio signal further used in the rapport estimator and so on.

Researchers at the Institute of Creative Technologies from the University of Southern California have designed a number of virtual characters for 2D displays. These characters are able to use natural language, perform appropriate gestures, show emotion and react to verbal and non-verbal stimuli, each of them being design for specific purposes. For example, SimSensei Kiosk Ellie is able to recognise and identify psychological distress such as PTSD, anxiety or depression from multiple signals with a specific aim of providing early support to military personnel [DeVault et al., 2014]. Two virtual human projects were designed to be used at the Boston Museum of Science: Ada and Grace [Swartout et al., 2010, Traum et al., 2012] -museum tours and Coach Mike [Lane et al., 2011] -museum staff. The museum tours' main role is to engage the middle school children in science and technology topics, while the museum staff help visitors use the 'Robot Park' exhibit. There are also a number of agents designed for training purposes including negotiation prototypes, game-based simulations for soldiers or simulations for interviewing bombing suspects [Kenny et al., 2007, Terada et al., 2021]. Part of this category are also the virtual patients [Kenny et al., 2009] that use realistic scenarios for both military and non-military issues, in order to train future clinicians in therapeutic interview skills; and Virtual Interactive Training Agent (VITA) [Burke et al., 2016] who is built

for practising job interviews and specially designed to reduce anxiety in young adults with ASD (Autism Spectrum Disorder).

A recent version of Virtual Human Toolkit has been described by Hartholt et al. highlighting a few case studies [Hartholt et al., 2020]. These include the toolkit usage during a seated VR application where the agent is a virtual therapist [Gordon et al., 2019], during a room-scale VR interaction with the aim of exploring the virtual area with an agent, and in MR using the HoloLens, where the agent took the role of a job interviewer [Hartholt et al., 2019].

FAtiMA is another toolkit for creating autonomous virtual characters that can evoke empathic response [Dias et al., 2014, Mascarenhas et al., 2022]. This toolkit's core is a computational model of emotion and decision-making that can be incorporated when building virtual characters. Mascarenhas et al. highlight different use cases of this toolkit: a single-player mobile game where players provide customer service to different customers with unique emotional profiles; serious games for practising social skills [Mascarenhas et al., 2018]; or a VR demo where users embody a police interviewer and interact with a VC suspect [Mascarenhas et al., 2022].

Most of these virtual characters are developed based on the SmartBody character animation system [Thiebaux et al., 2008] and uses the Virtual Human Toolkit architecture [Hartholt et al., 2013, Hartholt et al., 2022]. In order to respond to the user's multimodal inputs (actions, expression, nonverbal behaviour, and so on) the system tracks the user using the MultiSense framework [Stratou and Morency, 2017] as well as a number of other modules for head orientation, face tracking, gaze direction analysing expression and audio data (for further information please refer to [DeVault et al., 2014]).

Shortcomings of using VCs form 2D environments into IVEs. Immersive Virtual Environments aim to simulate reality such that people would respond naturally to the situations and events that happen within the virtual environment setting. Two illusions are linked to the user's presence: place and plausibility illusions [Slater, 2009].

- *place illusion* refers to the user's feeling that they are in the virtual environment, knowing they are not actually there
- *plausibility illusion* covers the objects'/people's response to the user's actions in the virtual environment and their behaviour based on the user's expectation

The user's feeling of presence decreases or is absent when any of these illusions are broken. This leads to users not reacting to the events happening in the virtual world, but to the ones from the world they are physically in.

Virtual characters' behaviour highly influences the user's plausibility illusion. If the VC is limited in sensing what is happening during the interaction with the user, their behavioural response might not be as expected by the user and hence, the plausibility illusion will drop. Humans have very good observational skills: they can easily spot when the VC's verbal and nonverbal behaviours

are asynchronous or not responsive, decreasing their plausibility illusion and hence, their level of presence [Neff and Pelachaud, 2017].

Although there is extensive research on virtual agents for 2D screens, not all of the contributions from these works are fully transferable and compatible with immersive virtual environments. In many cases, the virtual agents presented above rely on rich facial expressions, the users have a limited area to look at (the monitor) and there is a small amount of body movement that can be taken into account. IVEs allow for richer interactions where the user can immerse in an environment and make use of more low-level social signals as well as more abstract non-verbal behaviours such as social attitudes.

Social attitudes are very little studied for IVEs. We identified this gap and designed a study in VR to detect social attitudes (see Chapter 4). Further, in the next section, we cover the literature and state of art regarding social attitudes in IVEs.

2.2.3 Social attitudes

Detecting social attitudes

As seen in the previous section 2.2.1, there is extensive research on low-level non-verbal behaviours. There are however more complex social behaviours that are less researched and not as straightforward to detect. Examples of these social behaviours are social attitudes, which refer to an abstract activity or action taking place in social interactions, such as sympathy, affection, aggression, or social engagement. These social attitudes are too complex to be identified using a set of rules. However, people can identify them in human-human (or human-VC) interactions and can interpret complex non-verbal behaviours, even from a still image [Vinciarelli et al., 2011].

There is increasing literature on detecting different social attitudes such as dominance, agreement or engagement in interactions [Dermouche and Pelachaud, 2019a, Khaki et al., 2016, Bee et al., 2009]. These tackle the interaction from video recordings and could be applicable to VC on 2D displays. They make use of features such as prosodic information, gaze direction, turn taking or facial expressions (action units). Though these studies are influential contributions to the field, they are not applicable in IVEs. This is because not all user's features are traceable and because the interaction in virtual environments has more dimensions available. For instance, social interactions in IVEs allow for other nonverbal behaviours, such as proximity, that are distorted on a 2D screen.

Social engagement is an important aspect to consider during user-VC interactions. As with other social attitudes, the VC should adapt its behaviour with a change in the engagement level. This has been researched on many occasions [Gordon et al., 2016, Woolf et al., 2009, Bohus and Horvitz, 2014, Dhamija and Boulton, 2017]. They propose methods that tackle engagement in interactions, however, they disregard the user-VC interaction dynamics loop during the model training.

Dermouche and Pelachaud include this loop in their work, detecting the engagement from dialogue videos on a 5-level engagement scale [Dermouche and Pelachaud, 2019a]. They also assess the ML

models that use only one person's data. However, these models show lower performance than the one taking into account both people's data. They train the model on actions units (AUs), head rotation, gaze angle and the conversational state of the interaction. They report the AUs as the feature that has the highest contribution to the model's performance. When trained on this alone, the model's performance on detecting engagement is 98% improving to 99% when all features are used [Dermouche and Pelachaud, 2019a].

User-VC rapport

The rapport between the user and the VC has an essential role in their social interactions, especially when it comes to unstructured situations. For instance, in a general practitioner training, the doctor's nonverbal behaviours link to the overall rapport between themselves and the user. The doctors' nonverbal behaviours also influence the viewers' perception of their empathy. [Pan et al., 2018, Pan et al., 2016] In humans' social interactions, people adapt and adjust their verbal and nonverbal behaviours based on their partner's behaviour and the overall social interaction [Burgoon et al., 2006]. Work such as [Dermouche and Pelachaud, 2019b, Feng et al., 2017], takes into account all participants in that interaction to train an ML model to detect or generate different elements of social interactions; however, social elements (attitudes) between the user and the VC also influence the interaction dynamics.

Taking into account this aspect in human-VC interactions makes the VC's behaviour flexible and able to adapt to the scenario at hand. Being able to detect different interaction dynamics and attitudes between the user and VC could be used to develop behavioural models. These drive the VC in social interactions, its nonverbal behaviour being dependant on the rapport/empathy between the user(s) and the VC [Cafaro et al., 2016].

Training an ML model with the data from all participants in an interaction better illustrates the interaction, leading to more robust outcomes when compared to data from only one participant. This is because, during dyadic human-human social interactions, one person's behaviour highly influences the behaviour of the other person [Steed and Schroeder, 2015, Burgoon et al., 2006]. Moreover, humans behave differently depending on whom they are interacting with, their culture or their upbringing. People's behaviour also relates to whether they are by themselves or in someone else's presence [Schilbach et al., 2013]. Hence the behaviour learnt from one person is not as applicable to the one learned from multi-party interaction.

In summary, social attitudes are complex behaviours that are difficult to automatically detect. Social attitudes are also impacted by the rapport and the interaction dynamic loop between the participants in an interaction. Having covered the grounds on nonverbal behaviours for virtual characters, the next section introduces how the appearance of avatars influences social interactions. It focuses on the avatar appearance, and how the perception of using different types of avatars changes over time, it is influenced by the user-avatar resemblance and the task the users are performing.

2.3 Evaluating Avatars' Appearance

2.3.1 Avatars in immersive virtual environments

Avatars are virtual characters whose behaviours are driven in real time by the humans they represent. Representing humans' behaviours in virtual spaces has an important impact on the feeling of co-presence. Co-presence is defined by Salter et al. as the sense of being and acting with others in a virtual space [Slater et al., 2000]. Therefore, in most cases, accurately representing the user behaviour can increase its sense of being in the virtual environment and consequently the co-presence. While the current technology allows for various methods of reproducing the user's general whole-body movements, reproducing subtle facial gesturing or gaze behaviour is still a challenging task. When avatars and agents are designed for 2D displays, the users' facial behaviours can be recorded and analysed; this task is more troublesome when it comes to an immersive virtual environment, as the current head-mounted displays cover the upper part of the face. This implies challenges for both developing avatars and building autonomous agents for immersive virtual environments, and this is because the agents' behaviours are built to react to the user's subtle cues such as eye behaviour or facial expressions.

A few studies tackle gaze behaviour for immersive virtual environments. Models of gaze behaviour are covered in Section 2.2.1. Pejsa et al. look at the conversational footing or signalling who has a certain speaking role in a group conversation [Pejsa et al., 2017]. Examples of these roles include being a speaker, addressee, bystander, or overhearer. They propose models of gaze and spatial orientation for a virtual agent that can be used to signal specific footing configurations. Based on their evaluations, participants match the conversational roles signalled by the agent, but these results are observed only in immersive virtual environment settings, not on 2D displays (flat screens). This suggests that people tend to be more sensitive to agents' non-verbal cues in immersive virtual environments than on 2D interfaces. On the other side, Seele et al. look into how important is to have avatars with realistic gaze behaviour, that would better reflect the users [Seele et al., 2017]. They compare three gaze models with different levels of realism fidelity: a model with saccadic movements only, a simulation model that extends the saccadic model, and a user's real eye movements recorded with an eye tracking device. The results show that the participants perceived high-quality communication in all three different gaze models, however, the authors describe a number of possible drawbacks and future improvements in their experiment setup. They mention that the task is not very gaze-dependent and the sample size is relatively small (21 pairs divided into 3 groups for each condition). They also suggest that the novelty factor of the virtual environment technology might have prevented the observation of subtle eye behaviour- an aspect that can be overcome in the future with people getting more familiar with such setups.

Another aspect of avatars in virtual environments is their appearance. Users can be represented as different types of avatars and their looks might impact their interaction with other users or their own impression of themselves.

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 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]. Table 2.1 summarises the literature studying the effect of different types of avatars in IVEs and face-to-face. The review was performed by using the keywords ‘avatars’, ‘virtual characters’, ‘VR’, ‘AR’, ‘MR’, ‘appearance’, and ‘longitudinal’ on literature search engines such as Google Scholar and Scopus. The review also includes work found as a snowball effect from the relevant literature. The table summarises the field of comparing different avatars’ appearance in IVEs, covering relevant work on longitudinal studies, different task contexts, the familiarity between participants, style and resemblance of avatars. Further, we detail these factors and the importance of temporality in the field of social interactions in IVEs.

Avatar Style

Avatar style can significantly impact the way users experience and perceive VR 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. As realistic representations are both more difficult to create and may evoke the uncanny valley effect [Lugrin et al., 2015], cartoon-style avatars that are more stylised and simplified are also a common way to represent users.

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]. In some cases, no significant differences were found between cartoon and realistic avatars [Yoon et al., 2019, Garcia et al., 2021]; in other studies on the realism of appearance, contradictory findings have been reported. Some research shows participants preferring realistic avatars [Yuan et al., 2019, Pakanen et al., 2022] and reporting higher quality of experience while using them [De Simone et al., 2019]. Latoschik et al. found that participants in their study reported higher body ownership when using realistic avatars compared to wooden-block-person ones [Latoschik et al., 2017]. On the other hand, a realistic appearance can induce the uncanny valley effect. This is especially evident with respect to faces, which 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]. Lugrin et al. 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 [Lugrin et al., 2015]. They found that the use of realistic avatars led to a lower illusion of virtual body ownership.

In addition to appearance, personality can also play a role in users’ affinity towards a virtual character. Zibrek et al. found that the user’s affinity towards the VC was based on the VC’s appearance and personality and that realism in VC appearance can be a positive choice in VR

[Zibrek et al., 2018]. Overall, the appearance of an avatar can greatly impact users' experience in VR environments, and it is important to consider the effects of different avatar styles when designing VR systems.

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., 2020, 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]. In two different works, one of Pan and Steed, and the other of Smith and Neff, they 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 [Pan and Steed, 2017, Smith and Neff, 2018]. Similarly, in two surveys with 16 and 87 participants respectively, Pakanen et al. asked participants to rank their first, second, and third preferred avatar in VR and MR [Pakanen et al., 2022]. Participants had to choose from 36 pictures of avatars with altered representation styles and body types. The most preferred avatar for both VR and MR was the realistic full-body avatar, with the full-body hologram avatar (which had a lower alpha, 'see-through', effect) being the second most popular choice for MR.

However, this does not always hold. Herrera et al. 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 [Herrera et al., 2020]. 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 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. 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 Table 2.1

Moustafa and Steed 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 [Moustafa and Steed, 2018]. 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. had dyads of acquainted participants embody both customised cartoon avatars and personalised realistic avatars (via video stream) [De Simone et al., 2019]. 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.

2.3.2 Tasks and Environment Setting

Many studies have looked at the impact of avatar appearance on various tasks, such as playing games [yoon2019effect, moustafa2018longitudinal, khojasteh2021working, langa2022multiparty, pan2017impact, smith2018communication, herrera2020effect, aseeri2021influence, pakanen2022nice] 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 and Kirchner 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 [Nordin Forsberg and Kirchner, 2021]. 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 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 [Sun and Won, 2021]. 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.

2.3.3 Temporality in IVEs Communication

Research on longitudinal studies in IVEs can provide insights into user behaviour changes. Due to the repeated exposure to the immersive environment, the novelty of using the environment fades, making the outcome more generalisable in a real-world situation. However, there are not many longitudinal studies in IVEs as they tend to take more time and resources to conduct. For example, in a study by Bailenson & Lee, participants experienced less simulation sickness and had a stronger connection with their team over time, but they did not report significant changes in the level of presence and co-presence [Bailenson and Yee, 2006]. Moustafa & Steed 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 [Moustafa and Steed, 2018]. The VR environment did not allow for nonverbal behaviours or facial expressions, so initially, participants had difficulty interpreting social cues. However, over time, they learned to rely on other cues such as voice tone.

Khojasteh and Won conducted a longitudinal study in Facebook Spaces where participants in dyads met for 5 sessions and played games in VR [Khojasteh and Won, 2021]. 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. conducted a longitudinal study which compared customised avatars to generic avatars [Han et al., 2022]. 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). The results show 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.

Author	Env.	Group size	Familiarity	Task	Context	Avatar style	Avatar body	Resemblance	Longit.	Contributions
Yoon '19	MR, VR	48; dyads ¹	strangers	puzzle, furniture placement	casual	cartoon, realistic	full-body, h&h, upper-body	pre-defined	no	Avatar body matters; no difference between avatar styles; best performance with realistic full-body.
Zibrek '18	VR	1106; indiv.	strangers	30s listening task	casual	rendering styles	full-body	pre-defined	no	Appearance and personality influence VC's affinity; realism is a positive choice for VCs in VR
Moustafa '18	VR	17; diff ²	friends, families	activities in GearVR	casual	cartoon	upper-body	pre-defined	yes: 4-5x	VR group dynamics and emotional states similar to F2F
Lugrin '15	VR	30; indiv.	-	find \ touch targets in a forest environment	casual	block, realistic, robot	full-body	pre-defined	no	Lower illusion of virtual body ownership in realistic avatars
Khojasteh '21	VR	20; dyads	strangers	5 games	casual	cartoon	upper-body	customisable	yes: 5x	Adaptation to VR increases over time; new ways to communicate in VR
Seymour '19	2D, VR	32; group	strangers*	listening to conversation	casual; conference	cartoon, realistic	upper-body	customisable, personalised	no	Realistic preference over Cartoon appearance.

Latoschik '17	VR	21; indiv.	-	wave in a mirror	casual	realistic, wooden	full-body	pre-defined	no	Realistic appearance shows stronger body ownership illusion;
Han '22	VR	81; diff ²	mixed ³	school tasks	education	realistic	upper-body	customisable	yes: 8x	Personalised avatars show increased self-presence but decreased enjoyment
Langa '22	VR	32; diff ²	acquaintance	charades	casual	point-cloud	upper-body	personalised	no	Platform for low-cost holographic communications
Pan '17	F2F, VR	48; dyads	strangers	games; discussions;	casual	cartoon, controllers	full-body, hands	pre-defined	no	Embodying avatars is preferred for VR; results similar to F2F
Smith '18	F2F, VR	60; dyads	strangers	negotiation and furniture placement	casual	controllers, robot	full-body, hands	pre-defined	no	Full-body VR embodiment shows a higher level of social presence, similar to F2F
Herrera '20	VR	102; dyads	strangers	20 questions game	casual	cartoon	full-body, h&h, static	pre-defined	no	Head&hands avatars outperform full-bodied ones on social presence, self-presence, and interpersonal attraction
De Simone '19	F2F, VR	32 dyads	acquaintance	watching videos	casual	cartoon, video-avatar	upper-body	customisable, personalised	no	Similar quality of experience for Video-VR and F2F but lowest for cartoon; more movement and direct gaze in VR

Aseeri '21	VR	36; dyads ¹	strangers	conver- sation, survival items, cha- rades	casual	con- trollers, scanned, video- avatar	full- body, h&h	pre-de- fined	no	Video-avatars show high trust and co-presence (as do scanned ones); avatar preference order: video, scanned, no-avatar
Freiwald '21	VR	17; dyads	strangers	snowball fight game	casual	realistic, non-hu- man	full-body	pre-de- fined	no	No significant difference on the avatar's appearance but rather on the locomotion
Sakurai '21	VR	8; dyads ¹	strangers	twister	casual	cartoon, realistic	full-body	pre-de- fined	no	The avatar appearance affected the interpersonal cognition for males only
Nordin '21	VR	44; diff ²	mixed ³	conference; business meeting	profes- sional, confer- ence	realistic	full- body, upper- body	cus- tomis- able, person- alised	no	Meeting the avatars weren't restrictive; conference avatars were enablers and obstacles for the interaction
Garcia '21	VR	105- survey; 12-VR task	-	lecture- listening task	educa- tion	cartoon, realistic	upper- body	pre-de- fined	no	VR user study: no difference on the familiarity, engagement, trust, humanness, and learning
Pakanen '22	MR, VR	16; dyads	strangers	game	casual	cartoon, realistic	full- body, static, upper- body	pre-de- fined	no	Comparison of different avatar appearance styles and body; the preferred one for MR and VR was realistic full body

Collingwoode-Williams '21	VR	17-dyads ¹ ; 18-dyads	strangers	investment collaborative game	casual	realistic, controllers	upper-body, hands	pre-defined	no	Consistency improves trust in an equal social dynamic in IVEs; the use of confederate could shift social dynamics.
Sun '21	VR	152; dyads	strangers	brainstorming	casual	cube, realistic	full-body, upper-body	personalised	no	No difference in perceiving the emotional state from the cube and realistic avatars; a link between proximity and emotional states.
Bailenson '06	VR	9; trias	strangers	games, problem-solving tasks	casual	realistic	upper-body	personalised	yes: 15x	Less simulation sickness, stronger team connection, less direct gaze over time

TABLE 2.1: Review of studies on avatars in IVEs and 2D environments. *Env.*: Environment; *Longit.*: Longitudinal; *indiv.*: individual; *F2F*: face-to-face; *h&h*: head&hands. ¹Dyads consist of user and confederate. ²Different group sizes. ³Participants were a mix of strangers and acquaintances. To ease the table navigation, the following are colour coded: Environment, Avatar Style, Avatar Body, Resemblance and Longitudinal.

2.4 Industry Application and Ecological Validity

2.4.1 The game industry

Players' interaction in VR games

The entertainment industry uses virtual environments to build creative and engaging experiences for their users. Virtual Reality devices offer a virtual medium that enables richer input mechanisms compared to traditional video games, allowing for novel interactions and different gameplay. Hence, many game companies develop and adapt their games to this platform.

The mechanism through which players interact in traditional video games is different from playing a VR game. In non-VR video games, players actively interact using traditional controllers (mice, keyboards, and/or game controllers). In VR, the players' input tends to be more complex allowing for 3D interactions that are closer to the real-life ones. Along with using the buttons on the hand controllers, users can play a game with a diverse range of motions. They can use their limbs, head or their whole body as a form of input to drive the interaction, as they would do in their day-to-day life. Engaging with their whole body in these activities allows for more immersive interactions in VR games, especially social interactions typically seen in narrative games.

In everyday life, we, humans don't interact with others or go about our daily tasks via the controllers popular in traditional video games. IVEs are meant to simulate reality. Hence making use of traditional games' approaches in a virtual environment impacts the user's sense of presence. Approaches from traditional video games could highly influence the plausibility illusion, particularly when it comes to the user's interaction with agents (or non-player characters, NPCs).

VR games with non-VR implementations

During everyday social interactions, we show complex behaviours: verbal (e.g., speech) and non-verbal (e.g., gestures, head movements, eye gaze). It is highly challenging to account for these behaviours, in particular for nonverbal behaviours, in games and media in general. In most cases, this is due to hardware and software limitations. However, these drawback leads to not implementing verbal/nonverbal behaviours in non-VR games or having them expressed via button presses or mouse/joystick move (in games such as *Heavy Rain* (quanticdream.com/en/heavy-rain), or *L.A. Noire* (rockstargames.com/lanoire)). In narrative VR games or immersive media, verbal and non-verbal behaviours are key factors, strengthening the plausibility illusion and the sense of presence in VR settings.

Because in a VR environment users can move freely, the interaction in these games does not have to be restricted by the buttons on game controllers. Although there are many applications in IVEs that rely on triggering events based on button-pressing, it does not have to be this way. The user's large and diverse range of inputs can be manipulated to design interactions with VCs that are closer to the ones taking place in real life. This aspect helps maintain the user's plausibility illusion. Consequently, this means that the user's experience of interacting with a VC is as similar as possible to an interaction that happens face-to-face with a real person.

Using the classic game mechanics and the interaction methods from non-VR games in VR ones, often influences the players' presence, negatively impacting the players' experience and leading to increased simulation sickness [Christensen et al., 2018].

There are popular VR games that make good use of natural interactions, such as Beat Saber (beatsaber.com) or SuperHot VR (superhotgame.com/vr). Most of them are not centred on a story or the interaction with NPCs. Dance Central (dancecentral.com) is another popular VR game where players dance based on instructions, mimicking dance movements from VC instructors. Although there are many NPCs whom the users can interact with, the interaction itself is done through a virtual mobile phone. Because of this, controlling the game is much closer to natural interactions as many people are used to handling a mobile phone on daily basis.

VCs (or in games, NPCs) are a core element in most games, with players being able to interact with them (fight/get help from them) or even have a dialogue with them. The outcome of the interaction often leads directly to the next actions available, making the interaction itself part of the game mechanics.

The narrative genre in games is designed around the user's interaction and dialogue with NPCs. For example, games that fall into this category include *Heavy Rain* (quanticdream.com/en/heavy-rain), *The Walking Dead Series* (skybound.com/telltales-the-walking-dead-the-definitive-series), *L.A. Noire* (rockstargames.com/lanoire) or *Mass Effect* (ea.com/en-gb/games/mass-effect). Here the dialogue is mainly implemented by selecting pre-defined phrases from a list, using the mouse, keyboard or joysticks.

It is more difficult to develop narrative games in VR with natural social interactions. This is because the natural interactions with NPCs are more complex than the interactions in non-narrative settings (such as slicing cubes with lightsabers- in Beat Saber). Other VR games, such as Half-Life Alyx (half-life.com/en/alyx), implement ways of interacting with the environment that are very close to how people do in daily life. Being able to open doors by pushing them, manually reloading weapons, crawling and freely moving around, enhances users' feeling of presence. However, most of the games like this one, rely on core mechanics such as shooting or fighting, making them violent. Having these violent behaviours happen in VR can have a strong and profound effect on the players' emotions and behaviour [Bailenson and Beall, 2006, Yoon et al., 2019], thus excluding users less interested in violent or action-based games.

Passive and active interaction in VR applications

VR devices enable richer input mechanisms by accessing the users' body movements and hence, users are able to engage and express themselves more naturally. This unlocks the possibility of using body movements (actions that they would naturally perform in real life), as a game input, leading to a new input interaction. The traditional *active interaction* (where, for example, the players would 'actively' select the option by clicking a button) can be replaced by a *passive interaction* (where the player's non-deliberate, 'passive' actions are inputs for a game). This way the user can interact with the non-player character through their non-verbal behaviours, unlocking a VR-specific way of interacting in games, especially in narrative VR games. Through rule-based methods, different

specific behaviours can be detected, such as raising hands, crawling or archery, as the hands and head need to be in a certain position and location one from the other. This allows the system to detect those actions by applying certain rules. However, more complex behaviours such as social attitudes (e.g., sympathy, affection, aggression or social engagement) are very difficult to detect using fixed rules. These behaviours enhance these applications/games, offering a more complex interaction system, which is richer and closer to how humans interact in real life. Consequently, it better simulates reality, increasing the user's plausibility illusion.

2.4.2 Avatars in remote meetings

IVEs are widely used in the work sector in areas such as film-making, training, education, medicine, therapy, remote collaborations and so on [Baniyadi et al., 2020, Bellanca et al., 2019, Stavroulia et al., 2019]. In many of these areas, using virtual environments keeps the costs low while allowing for repetitions (e.g., training or therapy). The value of using it for other sections is in the co-location of people and shared space to complete tasks (e.g., film-making, remote collaborations). Performing remote meetings in these environments helps preserve the spatial dynamics and social behaviours such as gaze targets and proximity. In most activities in virtual environments, users embody avatars. They represent the user's position, activity and identity. They can range from geometric shapes (spheres/rectangles with hands) to human-resembling upper-body or full-body shapes in different styles. The avatar style could also range from cartoon-like to hyper-realistic. A concern for realistic avatars is that they may trigger a mismatch between high expectations and delivery of nonverbal behaviour (i.e., movement, gesticulation, facial expressions), leading to decreased user affinity and feelings of unease [Shin et al., 2019]. Cartoon-like styling, whether generic or customised, may also lead users to be anxious about the appropriateness of non-realistic representation in a work context [Bailenson and Beall, 2006].

The majority of the research on avatars in virtual environments is focused on presence, workload or trust [Yoon et al., 2019, Khojasteh and Won, 2021, Heidicker et al., 2017], with mixed results [Latoschik et al., 2017, Yuan et al., 2019]. These topics are common because participants do not know each other before the study. However, in the case of work meetings, it is likely that participants know each other beforehand. They have seen each other either face-to-face or in video calls, hence they know how the others should look like.

Moreover, study participants 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]. Real-life collaborative work in immersive environments involves users who know each other and interact regularly, trying to get real work done. Then, the communicative functionality of avatars is essential. Since the spatial audio common to most immersive environments provides a highly naturalistic vocal representation, it is nonverbal communicative functionality that is primarily at issue, such as the ability to identify one another and then recognise facial expressions and gestures [Burgoon et al., 2016], negotiate proxemics [Hall et al., 1968], and to trust that the avatars have authentic representations [Oh et al., 2018].

2.4.3 Ecological validity in IVEs

Ecological validity refers to the ability to generalise the outcomes of an experiment run in a lab into the real-world situations that are studied [Schmuckler, 2001]. It is of high value to design experiments with outcomes that can be generalised to natural real-world behaviour. In this section we will focus on the ecological validity of some of the previous work on each of the three main topics of this thesis (people, agents, avatars), and how these motivated the decisions behind designing the studies presented in this thesis.

People. In Section 2.1 we covered in detail the theory of grounding in communication focusing then on the gaze and turn-taking nonverbal cues. This helps us understand how social interactions work between people.

In lab studies, participants often are asked to perform a certain task in social interactions. For better experimental control, the tasks are often structured. For instance, participants could be asked to move a cube from one location to another [Nguyen et al., 2018] or to speak during their pre-defined turn only [Cañigüeral et al., 2021]. This kind of behaviour during the experiment leads to restrictive social interactions. Consequently, the behaviours from these rigid setups do not generalise to how social interactions carry out in real-world environments.

We were motivated by this shortcoming and we addressed it in Chapter 3 in collaboration with the Social Neuroscience group from the Institute of Cognitive Neuroscience at the University College London who proposed the dataset from this study. We present the work on understanding the social dynamics between *People's* interaction with a focus on gaze. The dataset used was multimodal, covering rich gaze, speech and the participant's video data. The Social Neuroscience group at the Institute of Cognitive Neuroscience designed the study and collected the data. The dataset consisted of dyads of participants performing two unstructured tasks. Using this data we took into account the synchronised participants' behaviour data to illustrate the conversational dynamics. These aspects increase the ecological validity and hence the generalisability of the outcomes.

Agents. One of the building blocks of creating agents in IVEs is sensing and recognising behaviour from the interaction in order to respond accordingly (see Figure 2.1). We detailed the state of the art in this field in Section 2.2. As extensive research has been done in generating nonverbal behaviour for VCs, often it was implemented using data from one person. For example, VC's nonverbal behaviour is generated using the audio feature of their own speech disregarding the other person's behaviour in the interaction [Haag and Shimodaira, 2016, Ferstl and McDonnell, 2018]. These models do not take into account the sensing and responding loop, hence the behaviour generated might be asynchronous and not in line with the interaction dynamics. Consequently, it might impact negatively the user's plausibility illusion and it does not generalise to a real-world situation.

There is a challenge in taking the user's behaviour into account while generating the VC's synchronised and believable verbal and nonverbal behaviours during social interactions. To address this challenge, different attempts generate nonverbal behaviours from data where the participants have to perform various pre-defined tasks. In most cases, these tasks are highly structured, leading to restrictive social interactions. Therefore, the users perform only a concrete set of actions [Nguyen et al., 2018]. The nonverbal generation model from this data works on similarly task-dependant scenarios, unable to generalise on other situations. Hence, the lack of generalisation lowers the chances of the model being used outside of the lab.

Our study on *Agents* addresses these points regarding ecological validity. Our model of recognising social attitudes takes into account the data from both the agent and the user in order to generate the outcome. This is crucial for the targeted application- a narrative VR game. Keeping the players' plausibility illusion stable is key when it comes to keeping the players enjoying the game. Even though for the data collection in our study the participants were given a certain task to perform, it was not restrictive, and we used it only as an objective (an example of a task is: "try gaining the VC's trust"). Further, it did not specifically control the participants' verbal and nonverbal behaviour, allowing the collection of unrestricted data from participants. Lastly, this work was developed and run in close collaboration with two game companies: Dream Reality Interactive and Maze Theory. The project was also funded by the InnovateUK body. This allowed designing the study in the best way possible to have the ML model reused in both companies' own professional work, outside of the lab settings. Maze Theory's narrative VR game *Peaky Blinders* was the target project for our ML model and we used the case study of recognising social engagement for a virtual character in this game. Furthermore, as the proposed ML pipeline is generalisable to recognise other social attitudes, Dream Reality Interaction worked on a prototype karaoke game making use of this pipeline.

Avatars. In Section 2.3 we present the avatar's evaluation in immersive media and traditional 2D environments. Table 2.1 gives an overall view of the state of the art and highlights the gaps in this area. A particular shortcoming in this field is the limited longitudinal studies. Taking into account the user's behaviour over time provides richer insights into how the behaviour changes and how people interact with others once the novelty effect wanes. The results from repeated usage of the system (i.e., meetings in IVEs) can be generalised to the natural behaviour in real-world interactions. The majority of studies on avatars for IVEs and 2D screens are not longitudinal, their data coming from one-off interactions (for example:[Yoon et al., 2019, Langa et al., 2022, Pakanen et al., 2022]; see Table 2.1). The task the participants perform has a big factor when evaluating the appearance of the avatars they embody and the overall experience. A large proportion of the studies on avatar appearance base their results on pre-defined and rigid tasks such as waving the hand in front of a mirror [Latoschik et al., 2017] or furniture placement [Yoon et al., 2019, Smith and Neff, 2018]. Although these tasks allow for interesting results, the outcomes are not as generalised to the real-world use of these systems. Tasks that are less limiting and more free-flow (i.e., conversations) have higher ecological validity.

We attempt to address these shortcomings in Chapter 5 where we present our work in collaboration with the Future of Work and Mixed Reality labs from Microsoft Research Cambridge. Our proposed study takes place over the 2–3 weeks recording around 10 sessions of work meetings per each group of co-workers. The meetings are not pre-defined. The participants have their usual meetings with their co-workers. This allows collecting data on how the *Avatars* appearance interacts with the tasks during a work meeting. Furthermore, the meeting takes place in the participant’s house or office, rather than in a lab. These factors (longitudinal data from real-work meetings performed in natural settings) strengthen the ecological validity of the work, assuring that the outcomes on avatars’ appearance are generalised outside of the study set-up.

2.5 Summary of literature review

The aim of this work is to advance the area of autonomous VCs in IVEs during social interactions. Building on previous research means taking into account the theory on social interactions, being aware of grounding, the least collaborative effort, the grounding costs and their link to non-verbal behaviours. Diving in more specific non-verbal behaviours, gaze and speech-turns are core to social interactions. Hence, it is crucial to understand the dynamics of these non-verbal behaviours and how they are detected, analysed, and generated. Social attitudes are very difficult to be described with straightforward rules. They are very useful to detect, being applicable in many fields, especially in the narrative VR games area. Finally, the appearance of VCs and avatars can impact social interactions and how other users perceive them. These are key in social interactions in the industry, in particular in collaborative remote team meetings in immersive virtual environments. All these aspects are the building blocks and contributions to moving forward in the field of social interactions in virtual environments.

After this overview of the literature and the current state of the art in this area, we present the initial work on gaze and speech. Hence, Chapter 3 shows the work on the low-level non-verbal behaviour dynamics in dyadic social interactions. This work is the initial step towards the understanding of non-verbal behaviours from face-to-face interactions, which acts as the base for building models and advancing the area of autonomous agents.

3

PEOPLE: Direct Gaze and the Frequency of Gaze Change

In this chapter¹, we analyse gaze and speech behaviours to gain insights into conversational dynamics between dyads, during an unstructured free-flow conversation. Nonverbal cues have multiple roles in social encounters, with gaze behaviours facilitating interactions and conversational flow. Using automatic analysis (rather than manual labelling), we investigate how the gaze behaviour of one person is related to how much the other person changes their gaze (frequency in gaze change) and what their gaze target is (direct gaze-DG or avert gaze-AG). Our results show that when one person is looked at, they change their gaze direction with a higher frequency compared to when they are not looked at. They also tend to maintain a direct gaze on the other person when they are not looked at. The outcomes of this work contribute to a more realistic gaze model for agents, by modelling more complex dynamics for virtual characters. This could be applied to a wide range of VR applications, such as soft skill training, language learning, and entertainment.

3.1 Introduction

When we interact with other people we use both verbal and nonverbal signals, not only to make ourselves understood but also to check if the message is received as we intended. Gaze behaviour is one of the nonverbal cues that facilitates interaction and the conversation flow. Its dynamics can be very different in live social interaction from, for instance, when watching a video [Cañigüeral

¹Results published in: Dobre, Georgiana Cristina, Marco Gillies, Patrick Falk, Jamie A. Ward, Antonia F. de C. Hamilton, and Xueni Pan. "Direct gaze triggers higher frequency of gaze change: An automatic analysis of dyads in unstructured conversation." *In Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 735-739. 2021 <https://doi.org/10.1145/3462244.3479962>

et al., 2021]. Gaze is tightly coordinated with other nonverbal behaviours such as speech [Kendon, 1967, Ho et al., 2015]. Gaze is an important part of social interaction with many different functions. Some of these functions are: regulating the conversational turns [Kendon, 1967], providing extra information in ambiguous situations [Macdonald and Tatler, 2013], giving insights about how people think or feel [Baron-Cohen et al., 1997], getting the other person's attention or approval [Efran and Broughton, 1966], signalling attractiveness, dominance and threat [Ellyson et al., 1981, Argyle and Dean, 1965, Emery, 2000].

During a conversation, the amount of time one person looks at the other person varies considerably [Kendon, 1967, Exline, 1963]. On average, listeners tend to give more DG to speakers, with these DGs being broken only by very short AG periods [Kendon, 1967]. However, the DG and AG of the speaker tend to be, on average, more equal in length. In other words, the speaker's AG is considerably longer than that of the listener. This led to our first hypothesis *H1*, such that listeners performs more DG than speakers, which also serves as a validation of our data and method.

H1: Listeners perform more DG than speakers.

Another factor in gaze is an approach/avoidance conflict. Argyle and Dean [Argyle and Dean, 1965] proposed that eye contact or mutual gaze (when both participants are gazing at each other) is, on one hand, actively sought in conversations for increasingly closeness and self-validation. On the other hand, there is a tendency to avoid excessive mutual gaze, as it can be overly intimate and arousing. This leads to a conflict that is normally resolved by reaching an equilibrium level of mutual gaze. This was the motivation for our second hypothesis *H2* when a participant is being looked at by their conversational partner (i.e., when their own gaze towards the partner would result in mutual gaze), they will be more actively managing the level of mutual gaze through their own gaze behaviour and will therefore switch between directed gaze (towards the partner) and averted (away) more often (*H2a*). They will also look less overall at the other person (lower overall mutual gaze), *H2b*.

H2a: When someone is being looked at (receiving DG), they would switch back and forth between performing DG and AG with a higher frequency compared to when they are not being looked at.

H2b: When someone is not being looked at (receiving AG), they would look more at the other person's face (DG) than somewhere else (AG).

Hence, the interaction dynamics between people are greatly influenced by the participants in the interaction. Looking at only one's behaviour gives only partial insights into nonverbal behaviours in interactions. Works such as [Dermouche and Pelachaud, 2019b, Ahuja et al., 2019, Feng et al., 2017] take into account data from all participants in that interaction to detect or generate different aspects of social interactions. Although it helps advance the field, a nonverbal behaviour model based solely on one's data leads to behaviour that is neither flexible nor contextual.

Understanding the gaze behaviour dynamics between two people can inform constructions of gaze and non-verbal behaviour models for conversational agents. This is particularly relevant now, as some Virtual Reality (VR) headsets come with gaze tracking capacity (e.g. VIVE Pro Eye), enabling a whole range of applications in gaming and social skills training, where gaze behaviours appear

to be similar to when taking place in real-world interactions [Sidenmark and Gellersen, 2019]. Importantly, the agent’s nonverbal behaviour has implications in maintaining the user’s plausibility illusion [Slater, 2009]. For instance, poor coordination between gestures and speech can make the agent be seen as nervous or not eloquent, while poorly timed gaze behaviour can disturb the conversation’s smooth flow [Neff and Pelachaud, 2017].

In previous works, researchers would either observe and label a social interaction live [Kendon and Cook, 1969] or they would video record the interaction for later analysis [Duncan, 1972]. Some of these studies look at a *structured task* between two participants [Freeth et al., 2013] when they have predefined actions such as going through a set of predefined questions. A benefit of these types of structured interactions is the ability to control better the conversational roles (speaker/listener). Although they do have important contributions to the field, the results from structured tasks are not always applicable in free-flow conversations, with clear limitations when used as a building block for nonverbal behaviour models used in autonomous virtual agents.

In *unstructured tasks*, participants are usually instructed to speak about a certain subject (free-flow conversations) or to speak with a confederate about a certain topic [Hessels et al., 2019]. These tasks are closer to how people interact every day and can capture different conversational dynamics between participants. Insights from studies with unstructured tasks could help create a nonverbal behaviour model for autonomous agents. Gaze behaviour, for example, is one of the social behaviours that has been well studied [Ho et al., 2015, Argyle and Ingham, 1972b, Kendon and Cook, 1969].

One major challenge in the analysis of unstructured conversation data is the annotation or labelling of the specific events within the recording, which are typically more time-consuming than structured ones. Although interesting results are emerging from these, it would be difficult to scale the manual annotations to large datasets. Also, it brings challenges when working with interactive autonomous agents, as the same manual data labelling needs to happen in real-time, making it not truly autonomous.

We aim to explore conversational dynamics between two people in a free-flow discussion that could be later integrated into a nonverbal model for an autonomous agent for real-time social interactions. We use automatic data annotation methods and considered the gaze targets of either looking at the other person’s face (direct gaze) or not looking at the other person’s gaze (avert gaze). We consider the following hypotheses:

- **H1: Listeners perform more DG than speakers.**
- **H2a: When someone is being looked at (receiving DG), they would switch back and forth between performing DG and AG with a higher frequency compared to when they are not being looked at.**
- **H2b: When someone is not being looked at (receiving AG), they would look more at the other person’s face (DG) than somewhere else (AG).**

This initial work aims to strengthen our understating of gaze dynamics. We plan to further include these findings in building nonverbal behaviour models for autonomous agents in virtual environments however, this is not covered in this chapter.

3.2 Dyadic multimodal dataset

The multimodal data was planned and recorded by the Social Neuroscience group at the Institute of Cognitive Neuroscience from University College London (antoniahamilton.com). The setup involved a room with two stools so that each pair of participants was facing each other at a distance of approximately 1.5 m (Fig. 3.1). A projector screen to the participants' side showed instructions, with pre-recorded audio cues played by a speaker. The researcher was separated from the participants with a curtain. They remained in the same room, but could not be seen nor did they interact with participants during the experiment. A video camera recorded the whole session. Each participant wore a lapel microphone. Their voices were registered on an audio file (left and right channels). Each of them also wore the PupilLab glasses (pupil-labs.com) that recorded their eye, gaze data, and a video stream of that person's view. Upper body motion capture was also recorded but excluded from this work.

There were 62 participants recruited from a local mailing list. They were paired up as 31 dyads given their availability. Participants acclimatised to the experimental set-up through a PupilLab glasses calibration session and a task of watching a short cartoon. We did not include these parts in the analysis. Next, they were engaged in three types of tasks: discussion, picture description and meal planning (recipe). There were five sessions in the following order: discussion 1, picture description 1, recipe, picture description 2 and discussion 2. The activity took on average one hour to complete. Here only discussion 1 & 2 and recipe were included as they both are unstructured tasks where participants were not told when to speak or listen. They were left to talk freely. During the discussion task, the participants talked about a three-minute short children's cartoon video that they previously watched together. The video had no words and it was about a drawn line creating obstacles for a character [Roberts, 2011]. The aim is to re-create a scene of remembering shared events with others. During the discussion task, both participants are asked to recall the events from the cartoon video. This task allowed participants to have a free-flow unstructured conversation, to discuss what happened in the video and help each other remember as many details as possible.

This task lasted two minutes, and took place on two occasions for each pair, resulting in a total of four minutes of dialogue for each dyad. In the recipe task, the participants spoke freely in order to plan a meal that uses ingredients both dislike. This task took approximately five minutes for each dyad.

The gaze target data was exported from the PupilLab software, and it can have low confidence when the eyes are closed (blinks) or when the target gaze can not be detected due to the eye shape, the participant's makeup, or if the PupilLab glasses were not well fitted. Out of the 31 dyads and 93 task datasets (three tasks per dyad), we removed 37 datasets as the overall gaze target confidence

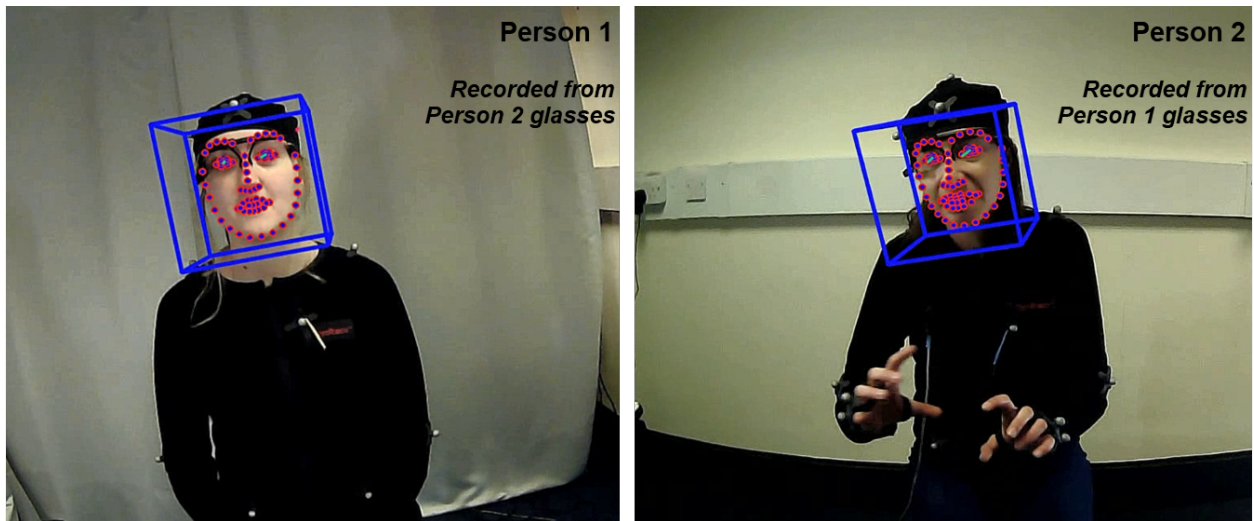


FIGURE 3.1: Setup for recording the dyadic data

Person 1 and 2 are seated in front of each other. The footage was recorded from a camera on each person’s PupilLab Glasses. The blue box and the facial landmarks were exported from OpenFace software and added to the original video.

was less than 65%. We consider 56 tasks from 23 dyads. There were 18 same-sex dyads and 5 mixed-sex dyads (41 female and 5 male).

Out of the total 56 datasets there were: 21 for discussion 1 ($D1$), 18 for discussion 2 ($D2$), and 17 for the recipe (R) task. A total of 164 minutes were recorded, with 78 from $D1$ and $D2$ combined, and 86 from R. Out of these 23 dyads, 11 had the speech recorded only from one lapel microphone due to a technical error. They were excluded from the speech-related results (i.e., H1). From those with full audio available, we considered all three tasks from 8 of the dyads, $D1$ from one dyad, R from one dyad, $D1$ and R from one dyad and finally both discussions ($D1$, $D2$) for one dyad. There are in total 10 recordings for $D1$ and R, and 9 for $D2$. This brings a total of 29 tasks and 88 minutes of data (50 from R and 38 from $D1&2$).

Data post-processing

We post-processed the data from the PupilLabs glasses and the audio files. Here, the term *audio* describes the sound that comes from a participant - it includes the speech but also laughter or backchannels.

From the PupilLabs software, we exported the gaze targets and the person’s view in video format. We used the video for getting the face location of the other person (the person they were looking at). To generate the face position data, we used OpenFace software [Zadeh et al., 2018]. From OpenFace we calculated a square to fit the participant’s face. However, as the returned values represented the face contour (excluding the forehead), we enlarged it with 10%, to capture the edge of the face. With the gaze target data for each participant and the face coordinates of the other person, we were able to detect the behaviour of looking or not looking at the other person’s face (DG/AG) by checking if the gaze target is inside the facial contour. The data was recorded at a

30Hz frames per second frequency. However, to limit the noise in the data, we scaled it down to 6 frames per second, combining each continuous 5 frames.

Each channel from the audio files was post-process by applying Google’s WebRTC Voice Activity Detector (webrtc.org/) via the python interface *pyvad* version 0.1.3 (pypi.org/project/pyvad/). The detector output was binary voiced or unvoiced data (value 1 or 0) per sample for each audio channel with a sample rate of 22050 Hz. Each channel represents one person from a given dyad. As the gaze data is represented with 6 frames per second (166ms) frequency, we used the same frequency for the audio data. We summed the values for each 166ms window: if the window was unvoiced, the resulted value was 0, whereas if the window was fully voiced, the resulted value for that window was 3675 (dividing the sample rate by six: $22050/6$). Hence, the outcome voice detection file had a frequency of 166ms, and each of these data points had a value between 0 and 3675.

Participants were in close proximity, hence the microphone from one person was recording some of the activity from the other person. We considered this when post-processing the voice detection files. Given person A with their microphone mA and person B with their microphone mB, in the ideal scenario, mA would record only A’s voice and mB only B’s voice. In reality, as A starts speaking (while B remains quiet), mA captures A’s speech, however, mB also captures some of this speech. In this situation, in the data from the voice detection file, the values from mA are higher than the values from mB (the voice detection files contain values between 0 and 3675, see above). Because of this, we compared the values from mA and mB by each data frame and marked as ‘speaking’ the person whose voice detection value is higher. If the value is equal, then both of them are marked as speaking. This is usually the case when both data points from mA and mB had the highest value (3675). After this second data post-processing, the voice detection file has binary values: 0 for listening and 1 for speaking.

This post-processing might also introduce very short speaking duration sections (less than one second) that are not from the person currently speaking but rather their microphone captured them from the other person’s speech. To tackle this issue, we filtered any sections of speech shorter than one second. This also removed some of the backchannels or laughter that appear in the audio as the voice activity detector does not account for them.

3.3 Data analysis and results

3.3.1 Gaze behaviour during conversational roles

Firstly, we analysed the data to validate the most common gaze behaviour recorded in previous literature [Kendon, 1967]. In line with our hypothesis, the speaker has a higher amount of AG behaviour (looking away from their partner’s face) while the listener has a higher DG (looking at their partner’s face). We split the data into two parts based on the conversation role label (speaking or listening). Then we calculated the percentage of which a participant is looking at their partner or is averting their gaze, for both parts. On average, the listener looked more at their partner (69%) while the speaker had a DG of (61%). The percentages differed based on the task. In *D1* and *D2*,

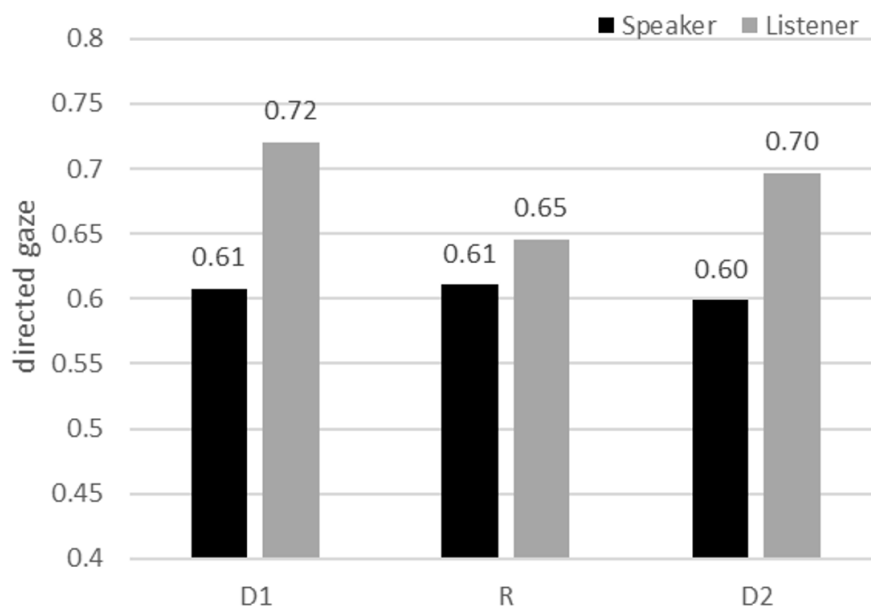


FIGURE 3.2: Direct Gaze percent speaking and listening

The y-axis shows the DG percent while being a speaker or a listener during each task. The tasks on the x-axis are in chronological order (*D* stands for discussion (1 and 2) and *R* stands for recipe)

the listener had 71% DG while the speaker only 60%. The difference is smaller in the Meal Planning (Recipe) task with 65% of direct gaze while listening and (61%) while speaking.

We performed a repeated measure two-way ANOVA with the conversational *role* (speaker and listener) and the *task* (D1, D2 and R) as factors. No *interaction* effect was found ($F(2, 7) = 4.148, p = 0.065, \eta^2 = 0.542$), and no effect was found for *Task* ($F(2, 7) = 0.238, p = .794, \eta^2 = 0.64$). However, there was an effect on *role* ($F(1, 8) = 71.024, p < .001, \eta^2 = 0.899$). As expected, the speaker performed significantly less DG, confirming **H1**. Figure 3.2 shows the values for each task and by role.

3.3.2 The effect of being looked at on own gaze

We were interested in the hypothesis that when someone is being looked at, they change their gaze differently compared to when they were not (*H2a*). Here we analyse how much they were changing their gaze behaviour per second. The gaze behaviour can be either *DG* (looking at the other person's face), or *AG* (looking away from the other person's face). Here we used all 56 tasks. We first separated the data into two datasets: when the participant is looked at (dataset L) and where they are not (dataset nL). We did this for each participant in the dyad. Next, we computed the sum of all the changes in gaze behaviour of the person being looked at (from dataset L) or not being looked at (from dataset nL). We then calculated how many seconds are in L and in nL. With these values, we calculated the frequency of gaze change per second by dividing the total seconds from the gaze change value (see Equations 3.1 & 3.2).

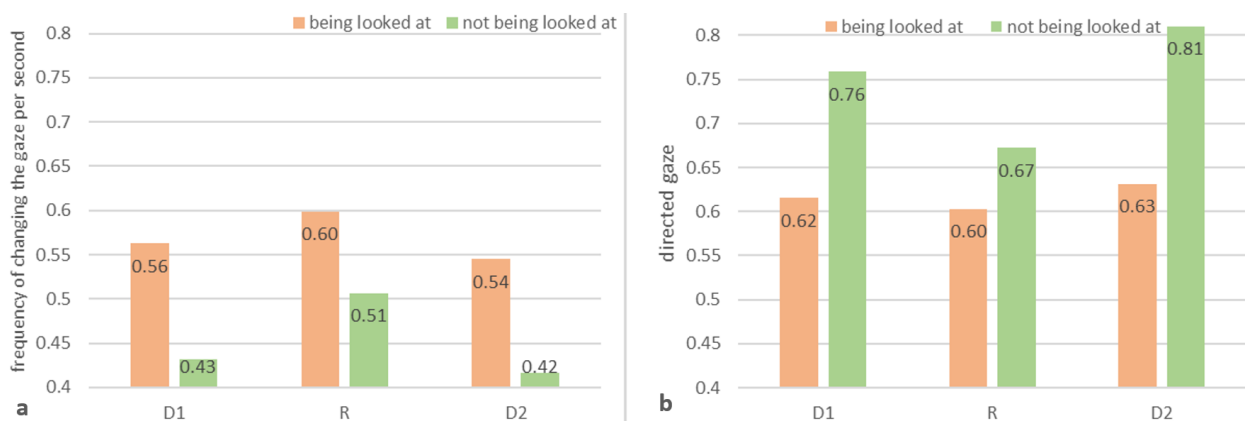


FIGURE 3.3: Gaze dynamics while being or while not being looked at

- a*: Gaze change frequency while being or while not being looked at during each task.
b: Direct Gaze percent while being or while not being looked at by task. The tasks are in chronological order (*D* stands for discussion (1 and 2) and *R* stands for recipe)

$$L_{gaze_change_fq} = \frac{\sum L_{gaze_change}}{\sum L_{duration}} \quad (3.1)$$

$$nL_{gaze_change_fq} = \frac{\sum nL_{gaze_change}}{\sum nL_{duration}} \quad (3.2)$$

Figure 3.3a shows the results by task, with the $nL_{gaze_change_fq}$ having lower value compared to $L_{gaze_change_fq}$. A repeated measure 3×2 two-way ANOVA (task and L: looked at/nL: not looked at). No *interaction* effect was found ($F(2, 15) = 1.001, p = 0.391, \eta^2 = 0.118$), so as for *Task* ($F(2, 15) = 1.590, p = .236, \eta^2 = 0.175$). However, there was an effect on **Looked at** ($F(1, 16) = 21.681, p < .001, \eta^2 = 0.575$). This confirms **H2a** that when being looked at, participants change their gaze pattern significantly more frequently compares to when they are not.

Given that participants were making fewer changes in gaze while they were not being looked at, we analysed behaviour during those periods, which led us to **H2b**. Hence, we calculated the percentage of DG of each person while being or not being looked at. We considered all 56 task datasets for this analysis.

As before, we first separated the data into two datasets: the part where the participant is looked at (dataset L) and the part where they are not looked at (dataset nL). We considered each participant in the dyad separately. Next, we summed the DG by the person being looked at (from dataset L) or not being looked at (from dataset nL). The percent is calculated by dividing the amount of direct gaze by the dataset size, either dataset L or dataset nL (see Equation 3.3 & 3.4).

$$L_{direct_gaze_percent} = \frac{\sum L_{direct_gaze}}{L_{size}} \quad (3.3)$$

$$nL_{direct_gaze_percent} = \frac{\sum nL_{direct_gaze}}{nL_{size}} \quad (3.4)$$

A repeated measure 3×2 two-way ANOVA was performed with two factors (task and gaze behaviour of their partner - L: looked at/nL: not looked at). No *interaction* effect was found ($F(2, 15) = 2.995, p = 0.080, \eta^2 = 0.285$), and no effect was found for *Task* ($F(2, 15) = 2.026, p = .166, \eta^2 = 0.213$). However, there was an effect on **Looked at** ($F(1, 16) = 28.091, p < .001, \eta^2 = 0.637$). This confirms **H2b** that when not being looked at, participants tended to perform more DG than when they were (see Figure 3.3b).

3.4 Limitations and discussion

A limitation of this study is our dataset which contains only dyads between strangers. The social dynamics cues between people can be different based on the familiarity of those participating in the conversation. Therefore, the results presented here might not generalise to different groups of people that are acquainted with each other from before the study. Further work is needed here to test whether our results stand if there is a level of familiarity between the conversation partners.

As with manual annotations, it is possible that our automatic data analysis process brings some degree of error. Hence, our H1 serves as a validation and was supported. We manually annotate a limited part of our data and the results were largely in agreement with our automatic analysis. We also used statistical tests to compensate for the noise. Further, with our automatic annotation method small errors can be compensated by the use of a large number of frames of data (higher frequency and longer time), also making it possible to scale to much larger datasets which would lead to better generalisation. Thus, we argue that the automated annotation method is one of the contributions of this study. However, a replication of this work using manual annotations could further strengthen our annotation method and results.

The Recipe task has a smaller effect for all three hypotheses. These differences in results could be explained by the task's nature. During the recipe task, the participants were asked to come up with a meal plan containing only foods both participants dislike. This led to silent periods where participants were thinking about the food they do not like, but also to more speech overlap, sections of laughter and backchannels. We believe that the smaller effect of this task is due to the nature of it. This is a great example of natural free-flow social interactions between people. This could take different shapes based on the main aim of the conversation and the common ground that gets built as the interaction progresses.

We used in this study multimodal data collected by the Institute of Cognitive Neuroscience from UCL. The whole dataset includes upper-body motion capture, speech and rich eye data. Our work from this chapter focused on speech turns and direct/averted gaze direction (looking or not looking at the other person in the dyad). More complex data could have been used, such as upper-body movement, and other low-level eye data (blinks/saccades). However, this was outside of the scope of this work. We believe that future work examining other social dynamics in free-flow conversation could bring important contributions to the field and could add to the insights presented in this chapter.

3.5 Summary

This work is the initial step towards the understanding of non-verbal behaviours from face-to-face interactions which act as the base for building models and advancing the area of autonomous agents. We analyse gaze and speech behaviour to gain insights into conversational dynamics between dyads, during an unstructured conversation. We used an automated method to annotate speech and gaze data from 56 unstructured tasks, from 46 participants. We found that people tended to have a higher frequency of gaze change (from averting to directing and vice versa) when they were being looked at compared to when they were not. During the times when the participants were being looked at, they were also directing their gaze to their partners more compared to when they were not. Alongside the proportions of gaze, we also looked at how it changes when being looked at (hence the use of gaze change frequency as a dependent variable). The outcomes are a direct contribution to understanding human interaction towards developing a diagnostic tool for neurological disorders such as autism and depression. Also, the work contributes to a more realistic gaze model for VR applications such as soft skill training, language learning, and entertainment, by modelling more complex dynamics for VCs.

After this insight into conversational dynamics, in Chapter 4 we move towards understanding and recognising higher-level and more complex nonverbal behaviours in social interactions. We cover social attitudes, with a case study on social engagement, and similarly to conversational dynamics, this work on social attitudes further contributes to building models for autonomous agents.

4

AGENTS: Immersive ML for Social Attitude Detection

In this chapter¹, we present our work on the development of a pipeline for training an ML model to detect social engagement. The pipeline features an immersive data collection and data annotation in VR for training an ML model to detect social engagement in a VC-user interaction. People can understand how human interaction unfolds and can pinpoint social attitudes such as showing interest or social engagement with a conversational partner. However, summarising this with a set of rules is difficult, as our judgement is sometimes subtle and subconscious. Hence, it is challenging to program NPCs to react towards social signals appropriately, which is important for immersive narrative games in VR. We collaborated with two game studios to develop an immersive machine learning pipeline for detecting social engagement. We collected data from participant-NPC interaction in VR, which was afterwards annotated in the same immersive environment. Game design is a creative process and it is vital to respect the designer's creative vision and judgement. We, therefore, view the annotation as a key part of the creative process. We trained a reinforcement learning algorithm (PPO) with imitation learning rewards using raw data (e.g., head position) and socially meaningful derived data (e.g. proxemics); we compared different ML configurations including pre-training and a temporal memory (Long Short-Term Memory algorithm - LSTM). The pre-training and LSTM configuration using derived data performed the best (84% F1-score, 83% accuracy). The models

¹Work published at Springer Virtual Reality: *Dobre, Georgiana Cristina, Marco Gillies, and Xueni Pan. "Immersive machine learning for social attitude detection in virtual reality narrative games." **Virtual Reality** 26, no. 4 2022: 1519-1538 <https://doi.org/10.1007/s10055-022-00644-4>* and at the Conference on Intelligent Virtual Agents: *Dobre, Georgiana Cristina, Marco Gillies, David C. Ranyard, Russell Harding, and Xueni Pan. "More than buttons on controllers: engaging social interactions in narrative VR games through social attitudes detection." **In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents**, pp. 1-8. 2022 <https://doi.org/10.1145/3514197.3551496>*

using raw data did not generalise. Given the pipeline’s results for social engagement detection, we generalise it for detecting human-defined social attitudes.

Overall, this work introduces an immersive ML pipeline for detecting social engagement in user-VC social interactions. We aim that the user-VC’s engagement information from the trained model will be used in the future by game creators as an add-on in their design pipelines. With it, they could enable the development of rich interactions for the VCs based on VC’s interactions with the players during the game. As these are our separate long-term goals, they are out of the scope of this chapter. Thus, this chapter covered the motivation for this project, the joint academia-industry work, the study design for the data collection, and the configuration, training and evaluation of different ML models.

4.1 Introduction

Complex human behaviours exhibited in everyday social interaction are hard to recognise automatically and therefore to use as a mechanic in video games. As a result, players often find themselves driving a social interaction in a video game by choosing what to do from a menu (see Section 2.4.1 for examples). In immersive virtual environments, this could break the plausibility illusion [Slater, 2009] and lead to break-in-presence [Slater and Steed, 2000], which takes the players back to the real-world and significantly reduces the level of immersion. In this chapter, we explore a novel pipeline in game design, combining ML and VR, with the aim to make social interactions in VR narrative games more engaging, immersive, and inclusive (in the sense that it will appeal to a broader audience than current video games).

VR devices could enable richer input mechanisms than that of traditional video games. In non-VR games, often players are limited to 2D user interfaces (keyboards, 2D game controllers). In VR, users can deploy a diverse range of motions in 3D: they can use their limbs, head, or their whole body as a form of input to drive the interaction, as they would do in their day-to-day life.

One of the most promising uses of body movement in VR is social interaction with VCs, or NPCs. In face-to-face interactions with people, we use our bodies extensively as non-verbal communication (colloquially called ‘Body Language’), including actions such as gaze (eye contact), gestures, posture and the use of personal space. VR opens the possibility to use these social cues as first-class elements of gameplay and thus creating much richer social experiences in games.

However, when the user input is more complex than button-pressing, it is a challenge to interpret its meaning in real time. Rule-based methods work well for detecting certain social behaviours when the hands and head need to be in a certain position and/or rotation (e.g., raising hands or looking at something). On the other hand, more complex social behaviours such as social attitudes (e.g. sympathy, affection, aggression or social engagement) are more difficult to detect using fixed rules, and we might even judge the same situation differently due to our personality and expectations.

Nevertheless, there are clear benefits to replacing traditional *explicit interactions* (selecting an option by clicking a button) with *implicit interactions* (social attitudes expressed via body language), where the player’s non-deliberate, *implicit* actions are inputs for a game [Schmidt, 2000]. This is in

particular important for maintaining the plausibility illusion in character-driven narrative games in VR where players could engage with NPCs naturally. Furthermore, being able to explore the use of social attitude with NPCs as a possible game mechanic (as opposed to, for instance, shooting NPCs with a gun) also makes games more inclusive and appealing to a wider demographic than current video games.

For this project, we worked closely with two immersive game studios (Dream Reality Interactive and Maze Theory) to create a pipeline for detecting such social attitudes that could be used as implicit interaction in a narrative VR game. The detected social attitude can then be used to trigger different behaviours in the NPC or in the game environment itself, influencing how the game continues. These triggers are to be decided by the game designers and game creators, based on how they envision the game and the gameplay. For instance, if a player is detected to be sympathetic to an NPC, they could gain a higher trust score from this NPC. Or if a player is perceived to be socially engaged with an NPC, the NPC could display animations that reflect a higher level of social engagement in return. The developed pipeline is meant to be integrated into the game studio's animation process in order to use the predictions to animate the NPC or the game environment. The pipeline is independent of any company-dedicated software as the set-up (Figure 4.2 part A, B) can be recreated and the algorithms used are available for implementation in other software. In this chapter, we focused on the social attitude detection part, the NPC's response to the user's social attitude in the game being out of our scope. Although we used the pipeline for the case of detecting social engagement, we argue that the pipeline can be generalised and used for other social attitudes such as sympathy, affection or aggression. For more details into the social attitude chosen, see the first challenge (gamer behaviour) in Section 4.2.

Game design is a creative process that involves the design of mechanics that guide players into certain desired behaviour patterns. While it is important that these behaviours in some way reflect players' natural inclinations, they are also defined by game designers who may want to guide players away from their more common patterns of behaviour. This is particularly true of the scenario we are studying in this chapter, as there was an explicit desire to guide players away from traditionally anti-social behaviour in narrative games towards pro-social interaction. Therefore, we view the work on social engagement detection as a creative interaction design process. Game designers should be in control of how the game, and characters in particular, respond to different actions in a player, just as, in traditional games, designers are in control of how the game responds to button presses. The definition of social engagement should not be viewed as an attempt to capture some objective measure (as might be done in traditional machine learning), but as a reflection of the game designer's creative judgement. The integration of machine learning into the creative process of game design and the foregrounding of creative judgement is one of the main contributions of this work.

The above-mentioned factors, that social attitudes are largely subconscious, that the behaviour is implicit and that this forms part of a creative project (a game), create a situation that we believe is relatively little studied. We are attempting to recognise a concept with no clear explicit definition. Social engagement, and certainly the behaviours associated with it are highly variable and contextual. If we were to attempt a definition it would be far higher level than the detail needed

for computational implementation. There is also no ground truth. Biometric measures might be used in some emotional contexts, but can only really distinguish low-level physiological states such as arousal, not high-level cognitive/emotional/social concepts like engagement. So we are dealing with a concept that can be defined only implicitly through human judgement. It might be possible to use player's own judgement of their feelings while interacting with a character, but these may not correspond well to their outward behaviour, it is perfectly possible for a person to be interested in what another is saying without outwardly displaying it, or conversely to outwardly appear highly engagement while inwardly feeling bored and thinking of other things (a fairly common human behaviour pattern). More importantly, the use of the player's own annotations would compromise the creative process. As described above how players interact with the game should be the result of a design process led by creative judgement. In this work we, therefore, treat the definition of social attitudes as a creative process driven by the judgement of a game designer. Social attitudes are, in this work, therefore concepts without explicit definition or ground truth and defined solely through expert creative judgement. This type of interaction design concept will be increasingly common as VR becomes a medium used by creative practitioners and which attempts to tap more complex and subtle aspects of human behaviour. Machine learning is particularly well suited to this task as it does not require an explicit definition at any point, simply a set of examples, which can be created through creative judgement. This is the key aim of this chapter.

4.2 Challenges and Contributions

Challenges

As we are collaborating with two game companies, we aim to develop a workflow which supports their creative design process and can be implemented into a production-ready VR game for the consumer market. At the beginning of the process, we identified our **three key challenges**:

1. **Gamer Behaviour**: this is part of a product that will be available on the market. Thus *it has to work for most gamers* (who will be paying for the game), which is very different from experimental studies with paid participants in the lab we were more accustomed to.
2. **Creative Process**: not only do we want to automatically detect a complex social attitude in real-time, but the expert annotator's judgement also has to be part of the creative process. In the game industry, Creative Directors define the artistic design of a game - we will need to include them as much as possible in this process.
3. **Market Reach**: the game has to be accessible for as many players as possible, meaning it will be developed cross-platform, considering the most commercially available headsets. This also means we are limited to the consumer market VR Headsets inputs (i.e., no access to eye, mouth, or EEG trackers) and software platforms that are compatible with major game consoles.

In order to tackle the first challenge, we need to better understand the **Gamer Behaviour**. After several in-depth discussions, we learnt from our industry partners that although players usually talk to other players in an online game, they almost never directly talk to NPCs. Thus, we needed

to create a scenario where the specific chosen social attitude could be present without the user speaking to the NPC. Through multiple brainstorming sessions (see Figure 4.1), we identified *social engagement* as an initial suitable social attitude to detect, as it could be present merely as a *listener* behaviour. It is also suitable for their current game in production, where the player has to gain the trust of various NPCs as part of their mission.

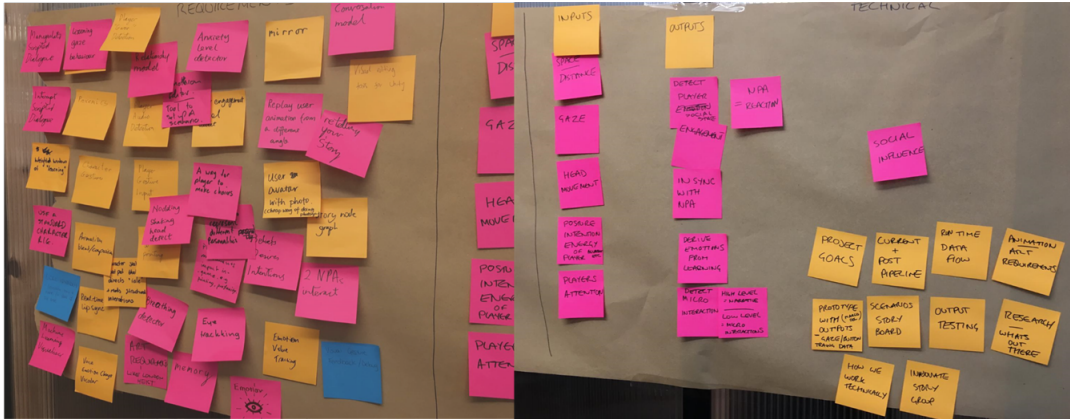


FIGURE 4.1: Notes from brainstorming sessions with collaborators from the game industry.

To address the challenge of detecting complex social behaviours and making it part of the **Creative Process** (in which the creative director could be involved as much as possible) we decided to explore the method of *imitation learning* where the ML algorithms could learn from their creative director. We taught an ML model to imitate how a human judges social interactions (Figure 4.2). We chose this because, as humans, we can easily detect the level of social engagement without always being able to verbally describe it, and different people might make different judgements for the same set up of behaviours due to their individual experiences. The ML model was trained on human annotations of the interaction between the user and the NPC in VR. The annotation also took place in VR, making it an immersive process (Figure 4.7). This enhances the annotator’s capability of observing the interaction from multiple angles and moving around the scene freely in the recorded interaction.

To make the game accessible to a broad market (**Market Reach**), we were limited to developing the pipeline with data captured from the player’s headset and hand controller. We were also platform restricted, having to design the data collection study and train the ML model within the Unity3D game engine (unity.com), using Unity ML-Agents. Further, we needed to use less complex ML models to reduce the computational cost, increasing the successful deployment and usage on all VR consumer devices, such as Oculus Quest, PSVR and PC-powered VR devices, in order for the game to work in real-time while maintaining frame rate required for running VR.

We chose an imitation learning approach, rather than, for example, a supervised classification in order for the method to fit more broadly within the framework of virtual agent behaviour used in industry. While supervised learning focuses on learning direct mappings between an input and output, reinforcement and imitation learning methods learn policies: probability distributions on the actions agents take in particular circumstances. A policy determines which actions should be

taken in a given state of the world and the agent. Thus, this is a direct driver of the agent’s behaviour. This focus on actions taken rather than mappings makes it well suited to modelling the behaviour of agents. Reinforcement learning is, for this reason, the most commonly used learning approach in the games industry [Shao et al., 2019]. This makes it appealing in our context for two reasons. Firstly, it is the most familiar approach in the games industry and is therefore more likely to be adopted. Secondly, it is more readily extensible to more complex agent behaviour models, which might not be the case for supervised learning. However, reinforcement learning per se is not suitable for this application, since it requires a well-defined measure of success or failure to use as a reward signal. In a standard game, the score or win/lose condition can be used, however, this does not apply to social interaction. Instead, we use imitation learning in which the reward signal is determined based on how well the agent’s behaviour matches a human demonstration.

Contributions

The main contribution of this work is the introduction of a creative director-focused pipeline for machine learning of social attitude detection. This pipeline provides **three principle novel contributions**:

1. Immersive experiment design
2. Immersive data annotation environment
3. ML model for implicit social attitudes detection.

First, we designed and conducted an experimental study of an immersive data collection process in which participants listened to an NPC’s monologue (prepared by professional writers from the national centre for immersive storytelling, StoryFutures Academy) in a VR environment closely resembling a real game social interaction. In three different VR stages, we gave participants either no instructions (*VR stage 1*), instructions that would very likely lead to socially engaged (*VR stage 2*) or socially disengaged behaviours (*VR stage 3*). Results from this experiment not only gave us useful insights into how players could behave in a VR game but also provided data to train our ML algorithms. Participants without instructions did not normally engage in social interactions, showing the benefit of providing realistic game tasks to guide behaviour during data capture (see Section 4.3.3).

Secondly, we developed an immersive environment where game designers could annotate the captured data, identifying instances of particular social attitudes. This VR environment placed the annotator in the same virtual space as the participant and the VC, enabling them to watch the interaction as if it were a real-life conversation. This allows, first, to make the most effective use of their social cognition, and second, to create an artist-friendly environment for data annotation, which is close to real gameplay experiences. The latter turns data annotation from a technical task to one that benefits from an interaction design skill.

Finally, with our pipeline, we were able to train an ML model to detect implicit social attitudes in VR interactions with 83% accuracy. Specifically, we used a reinforcement learning algorithm with imitation learning rewards from examples set by human experts. We report our comparison between

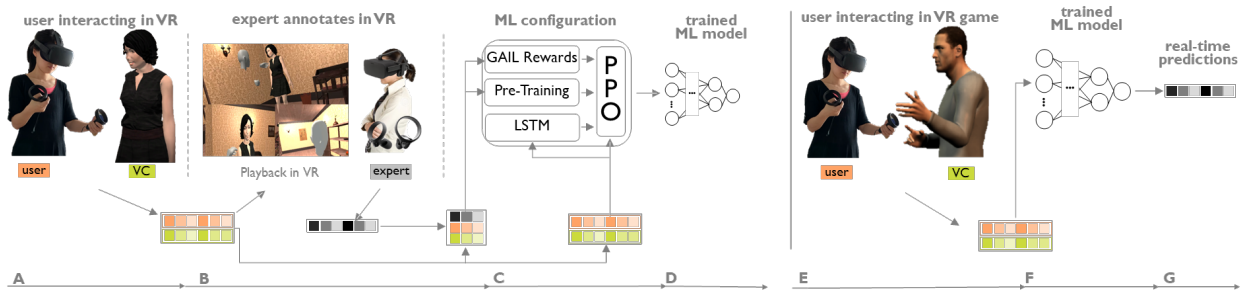


FIGURE 4.2: Pipeline for detecting human-defined social engagement

The pipeline includes an immersive data collection (user interaction (A) and expert annotating (B)) for training the machine learning model. This takes place by pre-training the model, creating Generative Adversarial Imitation Learning (GAIL) rewards for the reinforcement learning algorithm Proximal Policy Optimisation (PPO) that also uses a temporal memory called Long Short-Term Memory (LSTM) algorithm (C). This process exports a trained ML model (D). In a user-VC interaction (E), the trained model (F) detects in real time the human-defined social attitude (G) which could be used in different scenarios.

the model’s configurations and features that worked well. Our results show that pre-training the ML model improved the performance as did the use of temporal memory via an LSTM network. In addition, we propose several psychologically derived data features as inputs to the training which we show generalise better than raw features.

In the following, we describe how the pipeline is used for detecting social engagement in Section 4.3, where we also cover the experimental study with the data collection and the annotation process in VR. In Section 4.4 we describe how the model is trained, covering different input data and model configurations; then we present our results in Section 4.5. Section 4.6 explains how the pipeline could be generalised to be used as a detection tool for other social attitudes. We cover the limitations and discussion in Section 4.7 and conclude in Section 4.8.

4.3 Method: social engagement detection

Here, social engagement broadly refers to the social engagement one shows in social interactions linking it to the action of paying attention and showing interest. However, social engagement is a complex and subjective social attitude that is difficult to be described using concrete rules. Humans, on the other hand, have the ability to easily identify when social engagement takes place. Since this understanding is implicit, and we are designing a machine learning process based on creative judgements not on an objective definition, we do not formally define social engagement. Instead, the concept emerges implicitly from the annotator’s judgement of participants’ behaviour. In this Section, we describe how to detect social engagement between a user and a VC in an immersive VR scenario using the ML pipeline from Figure 4.2.

In the next part of this Section, we detail how we used the pipeline to collect data for detecting social engagement. We collected the data from users and then from the annotator. These processes happened separately but both in VR. We first describe the scenario we designed especially for

this data collection process (Section 4.3.1), then data collection with participants (Section 4.3.2), followed by how the annotation was done (Section 4.3.3) and finally, in Section 4.3.4 we present an overview of the human’s annotations and the questionnaire result (Section 4.3.5).

4.3.1 The scenario for data collection

We created an immersive and interactive VR scenario where users’ behaviour can be recorded. Specifically, users can interact with a VC (Figure 4.5 and Figure 4.4) created using Adobe Fuse Software (adobe.com/uk/products/fuse.html) and rigged using Mixamo (mixamo.com). This interaction took place in a room that we designed to reassemble a bedroom that will be used in the game, as suggested by the game company Maze Theory (Figure 4.3). The user can interact with, grab or change the location of the majority of objects in the room, for example, vanity box, birdcage, pillow, flower and vase, books, bin or chair, but not others, such as picture frame, poster, candle, rug, room divider.

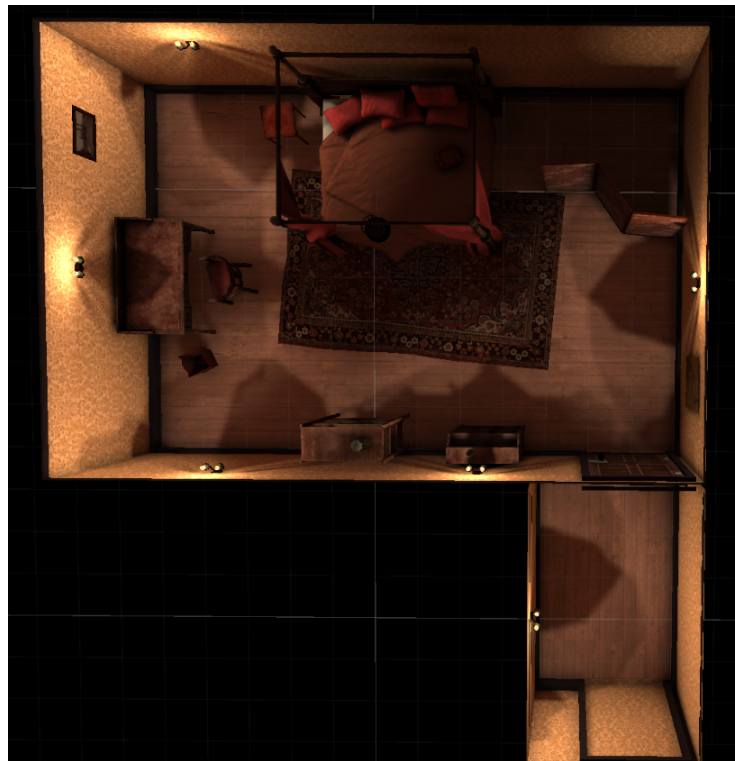


FIGURE 4.3: Representation of the environment for data collection

Top view of the environment. It contains the main bedroom and a small hallway. The user is able to freely move in the room and interact with most objects. It was created in close discussion and with design assistance from both game companies.

Virtual character implementation. During the interaction, the VC carried out a monologue about her family and her life. We collaborated with StoryFutures Academy, the national centre for immersive storytelling where professional writers wrote a captivating monologue. Table 4.1 illustrates part of the monologue; for the full monologue see Appendix C. While performing the



FIGURE 4.4: Frontal view of the virtual character

First person view of the VC that is used for the user interaction. It is shown with two different poses, looking at the user and looking into the distance. The VC is configured to fit the monologue scenario. It's modelled using Adobe Fuse Software and rigged using Mixamo.

monologue, the VC carried out animations for different behaviours as described in the monologue script. The VC performed generic animations using inverse kinematics to express specific behaviours. The VC maintained gaze and orientation towards the player for the whole duration of the monologue. There were short parts when the VC gazed and faced different objects in the room (see the full monologue in Appendix C). The VC's behaviour was scripted to collect data to train the ML model (see Section 4.4) and hence, it did not detect social engagement.

Study design. The study took place in VR and contained three stages based on the user instructions, which aimed to trigger both high and low social engagement behaviours in users. In the first VR stage (*S1*) the user was not given exact instructions. They were told to interact with the environment and the VC as they would do in gameplay, allowing us to study the range of different behaviours that participants would perform without prompting, gaining insights into the type of gameplay behaviours we could expect. In the second VR stage (*S2*), the user received instructions to try to gain the VC's trust, representing the kind of task players would be given in the game. This VR stage aims to record mostly high social engagement data. For the third and final VR stage (*S3*), the user received instructions to explore the room, representing a typical task that players would be familiar with from other games. The interaction from *S3* aimed to produce primarily low social engagement data. All tasks were designed based on feedback from our game developer partners from Dream Reality Interactive and Maze Theory to represent typical gameplay. For an example of the participant's behaviour in each part, see the video under the *Supplementary Information* on the paper publication website: <https://doi.org/10.1007/s10055-022-00644-4>.

The three VR stages took place in the same order for all users: *S1*, *S2*, and then *S3*. Since we are not comparing different stages, counterbalancing is not required. It was also not possible to



FIGURE 4.5: Example of users interacting with the VC

On the left, the user is listening while looking directly at the VC. In the middle, the user is patting the VC on the shoulder. Based on the user's questionnaire after the session, the user is interacting with the VC by trying "[...] to comfort her [the VC] by touching her shoulder when she was emotional [...]". In the image on the right, the user is interacting with objects in the environment, in this case, the user is swinging a birdcage.

counterbalance since doing Stages 2&3 before Stage 1 would prime participants' performance in Stage 1.

Ahead of *S1*, participants explored a training room that is similar to the room in the experiment, where they could interact with objects (open drawers/doors, grab objects) and move around the room. This extra step ensured the users were comfortable with the VR headset, navigation, and VR interaction techniques. All three VR stages and the training step took place in VR with an Oculus Rift Headset. It took about 5 minutes for each VR stage, resulting in each participant spending about 20 minutes in VR, with a small break between each VR stage when they filled in questionnaires.

4.3.2 User data collection in VR

There were in total 13 participants, 9 males and 4 females, aged between 20 and 46 years and an average of 32 years old. In terms of VR experience, 31% used VR less than 10 times, 38% more than 10 times but less than 50 and 31% more than 50 times. All participants voluntarily agreed to take part in the experiment and signed a consent form. The whole process was approved by the University's ethics board.

The data collection took place in two batches because of time and participants' availability restrictions. The first batch was with 6 participants and the second with 7. The only difference between the first and second batches is the VC's location and gaze direction (Figure 4.6). This difference was introduced to investigate the effect of the agents' gaze at various key objects, however, this did not give significant results and will not be discussed in this chapter. Nonetheless, this did allow for a more diverse dataset, where the VC had more than one location and variable head and body orientations.

Wistful monologue spoken with a sombre tone.
 VC: That's the only place we could laugh freely. The park with the rose finches. They've built apartments on it now. No longer can I ever go there. I wonder what happened to all the finches? Maybe they found a new home.
VC stares directly at the player again, her brow slightly crumpled.
 VC: Do you think they would have found a new home?
VC shakes her head briefly and her shoulders slump over a little bit.
 VC: No, they're like me, still looking for somewhere else to call home. I often imagine them happy[...]

TABLE 4.1: A snippet of VC's monologue.

The text in italic represents the scriptwriter's indications. As an interactive monologue, the user was directly addressed in sections such as *Do you think they would have found a new home?* The monologue was created by the StoryFutures Academy. For the monologue animation, see the video under the *Supplementary Information* on the paper publication website: <https://doi.org/10.1007/s10055-022-00644-4>.



FIGURE 4.6: Virtual Character's view in both data collection batches

The term *session* refers to each time the participant took part in the virtual scenario (regardless of the VR stage), hence, there are three sessions for each participant. There is a missing session from S3 in the second batch due to a software error, resulting in a total of 38 sessions, with 18 sessions from the first batch (6×3) and 20 from the second ($7 \times 3 - 1$). In total, the time spent in the VR environment by all participants is approximately 190 minutes ($38 \text{ sessions} \times 5 \text{ minutes per session}$).

We run the experiment in Unity3D and we collected data from both users and the VC. As described in Table 4.2, we recorded head, hands and root positions and rotations from the VC and the user. The root for the VC was situated in the hip, and in the head of the user. The root is not the same for the user and VC because the character model used was structured differently. Apart from that, we also collected the user's index and trigger buttons from the controller. They were using these buttons to grab objects in the scene. And lastly, we collected the user's headset velocity and angular velocity to capture the user's motion. We chose to collect the position and rotation data to record where the user and the VC are in the scene and where they are facing. The user's and VC's non-root (hands and head) information is relative to the root data as these elements are "children" of the root element in the Unity hierarchy. This data is then mapped in between -1 and 1 to meet the Unity ML-Agents recommended best practice (see Section 4.4.3). Because we used these values

straight from the trackers as they were available in the Unity3D engine, we refer to this dataset as *raw data*. Although there is no clear definition of raw data in the literature, in this chapter we use this notation to refer to the unaltered version of the data.

In total, over 108000 frames of data (multi-modal and at high frequency) were collected to train and evaluate the ML model (see Section 4.4 and Table 4.4).

Information Recorded	Data Type
User's head position	3D Vector
User's head rotation	Quaternion
User's left- and right-hand position	3D Vector
User's left- and right-hand rotation	Quaternion
User's main head anchor position	3D Vector
User's main head anchor rotation	Quaternion
User's left and right index and hand triggers	Float
User's Headset velocity and angular velocity	3D Vector
VC's head position	3D Vector
VC's head rotation	Quaternion
VC's left- and right-hand position	3D Vector
VC's left- and right-hand rotation	Quaternion
VC's main anchor (hip) and chest position	3D Vector
VC's main anchor (hip) and chest rotation	Quaternion

TABLE 4.2: Types of data recorded during the VC-user interaction

Data was recorded from participants and VC; 3D Vectors represent the X, Y and Z components in a vector data structure; The Quaternion represents the X, Y, Z, and W rotation components

4.3.3 Human annotations in VR

A human annotator watched a playback of the user interacting with the VC and annotated their interaction. As *social engagement* is a very subjective term and it has many definitions [Glas and Pelachaud, 2015], a human annotator marked the data without directly defining social engagement. In this case, the annotator implicitly defined social engagement by annotating it during the user-VC interactions. The annotator labelled the sessions' playbacks in random order. They did not know which VR stage or which user they were annotating.

To ensure the annotator had rich social interaction information, they could access the user's and the VC's camera view (showing their current viewpoint). This allowed the human annotator to have access to exactly what they were viewing at any time while being in the same place as the user and the VC. An example of this is seen in Figure 4.7 A. Different hand controller buttons ('A' and 'B') switched on/off the user's or the VC's camera view. The annotator marked the beginning of the high or low social engagement period, using the other hand controller buttons ('Y' and 'X' respectively, Figure 4.7 B). As they pressed 'Y' or 'X' the '-' or '+' signs coloured for 0.5 seconds with the corresponding colour (red or green). The '-' or '+' signs were on the annotator's (virtual) hand side.

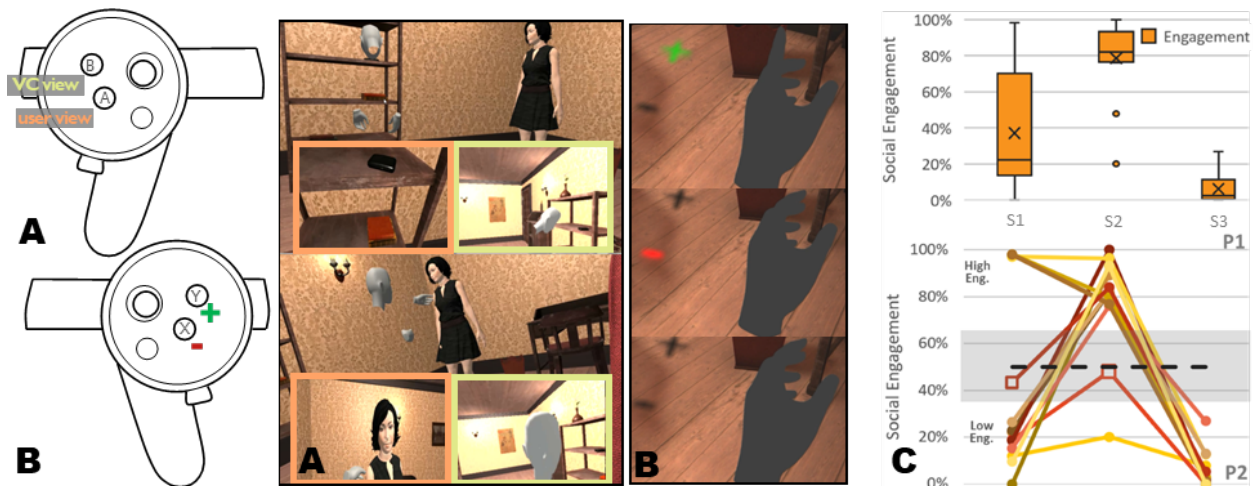


FIGURE 4.7: The expert's annotation process

Expert's annotations: **A** Controls mapping the camera view: the 'A' and 'B' buttons act as a switch to activate/deactivate the camera view from the user's or VC's perspective. **B** Controls mapping the social engagement level: 'X' and 'Y' records the current social engagement level rating, illustrated by colouring for 0.5 seconds the red '-' or the green '+' signs next to the (virtual) hand. **C** Engagement marked by the human annotator: **P1** shows the average (x) and median (line) percentage of all 38 sessions by the VR stage. These VR stages took place in the same chronological order (S1, S2, S3) for all participants. **P2** illustrates the average percentage of all sessions by the VR stage and by the participant. The back dotted line shows the 50% threshold that delimits the high from low social engagement sessions. Each other coloured line represents one participant. Sessions with a large mixture of low/high social engagement are positioned on a grey background and marked with a square.

The participant's avatar was represented in a simple way which showed only the head and hands in an abstract form (see Figure 4.5 & Figure 4.7 A). This was important as it removed features that were not accessible by the ML algorithm. Instead, the playback displayed only representations reconstructed from the data collected from the players. Therefore, it ensured that the annotator was not making judgements based on features that were inaccessible to the ML algorithm, and thus cannot be learned by it. Gillies et al. [Gillies et al., 2015] give examples of this problem. Annotators used video to annotate motion, but the learning algorithm used motion capture data. The result was that annotators (consciously or subconsciously) detected different behaviours based on features such as muscle tone or facial expression that were not available to the algorithm, which was therefore not able to learn to distinguish the movements. Although facial and voice information are relevant when it comes to social attitude detection (see Section 2.2.3), in this work we focused particularly on body gestures. This is because of the technical limitations imposed by the HMDs available in the current VR consumer market. We also decided not to include voice because each player could have very different background noise and different accents (making recognition challenging and unreliable), and we were informed by our game industry collaborators that gamers do not normally talk to NPCs (see the *Gamer Behaviour* and the *Accessibility* challenges in Section 4.1). Furthermore, the most important features from the literature (Section 2) such as gaze and body posture are strongly related to the feature we chose: head and body movements.

4.3.4 Annotations overview and validation

We computed the percentage of high/low social engagement labels between the user and the VC from each session by each VR stage. We calculated it as a percentage of all high (respectively low) social engagement frames over the total number of frames. As expected, when the users received instructions to behave with high or low social engagement (*S2* respectively *S3*), the users acted accordingly. For VR stage *S2*, the mean of high social engagement is 79%. Similarly, for VR stage *S3* the mean percentage of high social engagement is 6%. For the VR stage *S1*, however, the behaviour is mixed: most users showed low social engagement with some users displaying high social engagement. Figure 4.7 C-P1 shows these averaged levels over the three VR stages for all 38 sessions. In Figure 4.7 C-P2, these values are separated by each user, showing their behaviour in each session. The grey background colour highlights the sessions with a mix of high and low social engagement.

Most of the sessions have the expected engagement level (most of *S2* recording high social engagement and *S3* low social engagement level) with two exceptions for *S2*. Many users showed low social engagement when not given an instruction (*S1*). Although most of the averaged session's social engagement can be categorised as *low* or *high*, there are two sessions (from *S1* and *S2*) that are very close to the 50% threshold (marked with a black dotted line). These two sessions have a square marker in Figure 4.7 C-P2.

4.3.5 Questionnaire results

Participants answered a few questions after each VR stage. These questions were customised to the VR stage they just experienced. They could also leave some free comments about that stage.

VR stage 1. After *S1* (where they would hear the monologue for the first time without any instructions), they were asked to answer questions about the VC, such as: to list the family members the VC was talking about, the relationship the VC has with her family and how they think the VC was feeling; they were also allowed to write any comments about this stage.

Three participants wrote that they didn't listen to the VC and the other six that they stopped listening after a while; these participants had an incomplete or wrong list of family members or wrote that the VC's relationship with her family is '*loving, good memories*' (the VC was talking about the affair her mother had with her uncle and how her father didn't come to her mother's funeral). Based on the annotator's marking, these participants had either a low social engagement score, less than 15% (the ones who said they didn't listen) or between 19%-26% and one score of 43% for those mentioning they stopped listening after a while. The remaining four participants were able to answer the question about family members correctly, or almost correctly (two of them missed the mother, and one even mentioned the finch- the bird that the VC was talking about). In the general comments part, those participants also wrote about the way they perceived the monologue and what they think of the VC. Based on the human annotator, these participants had over 97% social engagement scores.

VR stage 2. Here, we instructed the participants to gain the VC’s trust. After this VR stage, they were asked how well they performed at gaining the trust and also to write any comments regarding this VR stage. The majority of them (11 out of 13) described what they tried to do to gain the VC’s trust. They said they listened, tried to be empathetic, nodded when appropriate and ‘*stopped messing around*’ because ‘*If I start looking through drawers and cupboards, I think she would be more suspicious of me.*’. These participants got over 76% high social engagement score based on the annotator’s marking. One of the remaining two participants said they were ‘*expecting some "helpers" to point out what you can or can’t do to a character*’ and that they could not earn her trust by themselves. This participant’s score of social engagement was 48%. The last participant wrote that they did not interact with the VC at all which reflects their low score of 20% given by the human’s annotations.

VR stage 3. After the last VR stage, where the instruction was to explore the room and remember as many objects as possible, the participants were asked to list all items they recall and to write any comments they have about this VR stage. All of them described how they explored the room and how that felt like: for some, it felt more immersive than the previous sessions, for others it was the opposite: ‘*having full control on exploring I lost a bit of immersion as I was behaving as I wouldn’t do in the real world*’. Others mentioned that the VC did not comment on them exploring the room (the VC having the same monologue as in the first sessions) or that they found ‘*it more interesting to interact with the objects while she speaks about them, (looking at the birdcage when she talks about the finches)*’. All participants reported a high number of items (from 9 to 17, with an average of 14), while in the room there were 23 items. As expected, the annotator gave low social engagement scores to all participants in this VR stage, as can be seen in Figure 4.7 C, P1 and P2.

In summary, the results for *S1* show that participants had a range of different behaviours when they were not prompted with a particular task, but with the majority biased towards low engagement. *S2* and *S3* were successful in generating the desired behaviour, using realistic gameplay tasks. This shows the benefit of giving data capture participants tasks to implicitly guide their behaviour (though the inclusion of unprompted behaviour could still be useful to identify unexpected behaviour patterns).

4.4 Training the detection component

We trained the model using imitation learning with the Unity ML-Agents platform (v0.11) and their main reinforcement learning algorithm Proximal Policy Optimization (PPO).

In this Section we explain the algorithms used (Section 4.4.1), then in Section 4.4.2 we cover the ML configuration, followed by what input data we considered (Section 4.4.3) and ending with Section 4.4.4, the ML implementation.

4.4.1 ML algorithms

We proposed different model structures, including pre-training with recorded data, and adding temporal memory through a recurrent neural network (Long Short-Term Memory: LSTM). Below

we present a brief description of each model.

PPO [Schulman et al., 2017] is a Reinforcement Learning algorithm and the idea behind these algorithms originated from behavioural psychology. It refers to an agent that changes its behaviour to maximise a reward function. The goal of a reinforcement learning algorithm is to develop a policy. This policy maps states to probabilities of selecting a certain action, with the aim of maximising the expected reward. More specifically, PPO trains a stochastic control policy where the agent learns its behaviour from experience without prior information on the environment or the task. Here, *stochastic* refers to having a probability distribution associated with all actions from each state.

To ensure that the model has a good starting point for optimisation, we pre-train the model using behaviour cloning (a simpler imitation learning model than GAIL, described below). This uses the training examples to find a good initial set of weights for the neural network for the full training algorithm.

GAIL [Ho and Ermon, 2016] is an approach to imitation learning, learning to execute a task by imitating human performance. It does this by generating reward signals from a human performance that are used to train the PPO reinforcement learning algorithm. It relies on data usually provided via human (or expert) demonstrations to learn a policy that behaves similarly to the human. The algorithm compares state-action pairs (each input and current circumstance with the corresponding response) from expert data against state-action pairs generated using the policy. At the same time, a classifier trains to differentiate expert data from the generated one. Thus, the policy develops to generate data that the classifier would mistake for the expert data.

Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] is a recurrent neural network. It learns a series of events with time order that have long time intervals. Based on this, it automatically determines the optimal time lags (time between two consecutive events), used for the next prediction. Its neural network is composed of one input layer, one output layer, and one recurrent hidden layer. The recurrent layer contains a memory block structure that memorises the temporal state and controls the information flow. It learns how past actions unfolded, thus knowing when to incorporate or drop past events and take future decisions.

4.4.2 Proposed ML configurations

PPO provides positive rewards for performing the desired behaviour and negative ones for the non-desired behaviours. In this case, to mimic an imitation learning scenario, the rewards are calculated using GAIL. These rewards show the performance of the action the model took and influence future actions.

We hypothesise that both pre-training it and adding a temporal memory (through LSTM) would improve the PPO's performance. The behaviour learned from pre-training influences the action taken by PPO, at the same time, PPO's policy attempts to maximise the reward. The temporal memory takes into account past actions, hence the algorithm considers past behaviour and current actions when deciding what to do next (what action to take). We hypothesise this because the behaviour that needs to be learnt is complex and temporal. We compare these models with those

Description	Datatype
Distance between the user and VC, based on Hall’s personal space [Hall, 1966], value mapped between 0 and 1	float
User’s facing direction: the angle between VC’s head rotation and user’s head rotations divided by 180	float
Interaction with objects: data from the controllers’ trigger (the trigger allows objects interaction)	float
User’s headset velocity	3D vector
User’s headset angular velocity	3D vector

TABLE 4.3: Socially meaningful derived data

These data are calculated based on the raw data detailed in Section 4.3.2 Table 4.2

without (using random initialisation instead of pre-training and a standard feedforward network instead of LSTM).

4.4.3 Input data for proposed ML training

There is strong evidence in the literature that certain behaviour aspects (such as body posture or gaze) are linked to social engagement [Mota and Picard, 2003, Sanghvi et al., 2011]. Based on this, we trained the model with psychologically-based features, such as the user’s facing direction, distance from the VC, interaction with other objects and their velocity and angular velocity (as shown in Table 4.3). Since these are calculated from the raw data we collected, we call them derived data.

The user’s distance from the VC is calculated based on Hall’s personal space [Hall, 1966]. We calibrate the virtual space units using the average human’s height of 1.65m [Max Roser and Ritchie, 2013] and mapped it to the user’s height in the virtual space units from the VR headset. Hall’s personal space has three different space layers: intimate (0.4m), personal (1.1m) and social space (3.6m). From these three, we use the intimate and social spaces as the lower and higher boundary. We calculate these thresholds from the VR headset height, which we assume represents the average person’s height (1.65m). Therefore, the 0.4m intimate space threshold is calculated from $\text{height}/4 = 0.4\text{m}$; and the 3.6m social space threshold from $\text{height}/0.45 = 3.6\text{m}$. The values are then mapped between 0 and 1. Thus, 0 is the further away from the VC: the maximum and above of *social space* and 1 is the closest to the VC: minimum of *intimate space* and below.

The use of derived data features inspired by the psychology of social interaction has the potential to improve ML performance. However, recent trends in Deep Learning have shown that deep neural networks are able to learn effective representations directly from raw data [Bengio et al., 2013]. In our evaluation, we, therefore, compare models trained on derived data with those trained directly on raw data. This comparison was done by training the best-performing model configuration on both raw data and derived data, detailed in Table 4.2. Both models (derived and raw data) use human annotations as ground truth data, and their output is a discrete binary value. The discrete value shows the current user’s social engagement at each frame. It can have a value of 1, for the user’s high social engagement, or -1 , for the user’s low social engagement.

	Sessions	Dataframes	High ann. %	Low ann. %
S1	13	37226	37.1	62.9
S2	13	37288	78.5	21.5
S3	12	33931	6.1	93.9
Total	38	108445	41.5	58.5

TABLE 4.4: The dataset used for the ML model training

Proportions of **High** and **Low** social engagement annotated data used for all three VR stages for training the model. In this table, *Ann.* is short for *Annotations* and *S1 – 3* for each VR stage.

These two options aim to mimic the human’s ratings of low/high social engagement during the annotation. To label the data, the human presses buttons for high or low social engagement behaviour; the data in between the button presses represents the most recent pressed value. For instance, if the annotator marks *high* at time t_i , *low* and time t_{i+1} , and then *high* again at t_{i+2} , all timeframes from t_i to t_{i+1} are labelled high social engagement, all timeframes from $t_i + 1$ to t_{i+2} are low social engagement and everything from t_{i+2} until the next button pressed is high again. This way, the annotations are in the same format and frequency as the model’s output, generating a value for each frame. The data has a frequency of 9 to 10 frames per second. We decided on this frequency as it has been used in the literature for low-level and subtle behaviour such as fast head nods [Hale et al., 2020].

The dataset for training with both raw and derived input data are detailed in Table 4.4, also showing the percent of the annotated low and high social engagement for all three VR stages. There are fewer dataframes in the last VR stage (*S3*) as there is a missing session due to a software error. The missing session doesn’t unbalance the dataset because the expected social engagement from that VR stage is low engagement, and there are already more than half low social engagement sessions from *S1* (see Table 4.4 and Figure 4.7 C-P2.)

4.4.4 Implementation

We analysed two additions to the PPO ML structure: pre-training and a temporal memory via LSTM. Therefore, we compare the PPO algorithm implemented with different configurations: with and without pre-training it, and with and without LSTM.

We randomised the dataset sessions and divided them into three folds of 13, 13 and 12 sessions each, for a 3-fold cross-validation; the training data consists of two folds while the remaining one represents the evaluation data. The hyper-parameters are tuned for both models with derived and raw input data (see 4.5). The hyper-parameters corresponding to *LSTM* and *pre-training* are dropped for the training configurations where these models are not used (in PPO+GAIL+LSTM, PPO+GAIL+PreTrain or PPO+GAIL).

Model	Hyper-parameter	Value (D)	Value (R)
PPO	batch_size	64	64
PPO	beta	5.0e-4	5.0e-4
PPO	num_epoch	5	3
PPO	buffer_size	2048	2048
PPO	epsilon	0.1	0.1
PPO	batches_per_epoch	10	10
PPO	hidden_units*	128	512
PPO	lambda	.99	.99
PPO	learning_rate	1.0e-4	1.0e-4
PPO	normalize	false	false
PPO	max_steps	75000	150000
PPO	time_horizon	128	128
PPO	num_layers	2	2
PPO	summary_freq	1000	1000
LSTM	use_recurrent	true	true
LSTM	memory_size*	256	384
LSTM	sequence_lenght	64	64
LSTM	learning_rate_schedule	linear	linear
pre-train	pretraining/strenght	.5	.5
pre-train	pretraining/steps	1000	1000
GAIL	gail/strenght	.5	.5
GAIL	gail/gamma	.9	.9
GAIL	gail/use_actions	true	true

TABLE 4.5: Hyper-parameters and their values for the ML models

These are used to train the PPO algorithm. (*D*) stands for *Derived*, while (*R*) stands for *Raw*, referring to the dataset used for training the different models. *The derived data was used for all four configurations, with the only modification of *memory_size=128* and *hidden_units=256* for PPO+GAIL, PPO+GAIL+LSTM and PPO+GAIL+PreTrain. These modifications took place as a result of hyper-parameters tuning, training the models with these hyper-parameters increasing accuracy and F1-score performance.

4.5 Results

In this Section, we present the results of the presented pipeline. Section 4.5.1 covers how the model’s prediction data is post-processed to match the format of the ground truth data. In Section 4.5.2 we present the results of the models trained with derived features while in Section 4.5.3 we provide the comparison of the model trained with derived features and the model that uses the raw dataset.

4.5.1 Data post-processing

We post-process the data on two different occasions: (1) we smooth out the model’s predictions data to remove noise and (2) we averaged the model’s predictions and the ground truth data from a 1-second section. The latter process returns one value for each section, which will be used to compare the model’s predictions to the ground truth data. We detail the post-processing actions in the remainder of this Section.

First data post-processing

We describe how the human annotates the ground truth data in Section 4.3.3. Briefly, the annotator marks only the change in the social engagement (from low to high or from high to low engagement), thus the ground truth data contains large blocks of either low or high social engagement data. The ML model outputs the predictions in a different way: it predicts a social engagement value at each frame. In many cases, this can result in noisy output, with regions of low social engagement containing a few frames of high social engagement (or vice versa). For instance, if we take a segment of length 10 dataframes (approx. one second), it can contain a majority of high engagement values, say 8, the remaining 2 being low engagement values. If we would compare frame-by-frame, the 2 low social engagement values are in minority in that window, and they can be seen as noise. When evaluated against the ground truth data, the 2 dataframes would appear as false negatives if the whole window would have high social engagement values.

Because of this difference between how the annotator created the ground truth data and how the models output the predictions, we post-process only the model's output data to remove the noise.

We smooth it out by applying a rolling window of 0.5 seconds on the model's outcome (Equation 4.1). This results in a float value; because it is not compatible with the ground truth data (integer datatype), we average the result to 1 if the rolling window result is higher than 0 and to -1 otherwise (Equation 4.4). The post-processing is further explained below:

Generically, a rolling window can be represented as:

$$W_i^{j,h} = \{x_{i-j}, \dots, x_i, \dots, x_{i+h}\} \quad j, h \in \mathbb{N}. \quad (4.1)$$

The number of samples in $W_i^{j,h}$ being: $|W_i^{j,h}| = j + h + 1$. For a 0.5s window size on a 10fps frequency, the number of samples is 5, hence the values for j and h could be 2 and 2 respectively, creating a symmetric window centred in x_i .

Given X containing all dataframes from a session, such as:

$$X = \{x_1, x_2, \dots, x_n\} \quad (4.2)$$

a window can be represented as:

$$W_i^{2,2} = \{x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}\}, \quad i \in [3, |X| - 2] \quad (4.3)$$

Then, the value for a dataframe (x_i) from X is:

$$x_i = \frac{1}{|W_i^{2,2}|} \sum_{x_m \in W_i^{2,2}} x_m \begin{cases} -1, & x_i \leq 0 \\ 1, & x_i > 0 \end{cases} \quad (4.4)$$

Second data post-processing

Although the ML model outputs a value at each time frame, the social attitude is very unlikely to switch from one state to another and then back to the initial state in a very short period of time (one-tenth of a second). Similarly, in other studies, the authors consider certain time sections. For instance, Yu et al. manually divide the conversation in utterances and use those for prediction [Yu et al., 2004] and Bohus and Horvitz consider a 5-seconds section for forecasting disengagement [Bohus and Horvitz, 2014].

We take a similar approach and average both ground-truth data and the predicted data over a time of one second. With this, we compare the model’s output to the ground-truth data and calculate the performance metrics.

We calculate the mean value from each time section, then we round the result to use: 1 if the mean is greater than .49; 0 if the mean is in between $-.49$ and $.49$; and -1 if the mean is smaller than $-.49$. Thus, given the time section S_t :

$$S_t = (x_t, x_{t+1}], \quad t \in \mathbb{N}, \quad t \in [0, T] \quad (4.5)$$

and T is the session length in seconds, then the value V of each section is:

$$V_{section} = \frac{1}{|S_t|} \sum_{x_m \in S_t} x_m \begin{cases} 1, & V_{section} > .49 \\ 0, & -.49 \leq V_{Section} \leq .49 \\ -1, & V_{section} < -.49 \end{cases} \quad (4.6)$$

The results have three categories: 1 for *High* social engagement, 0 for *Mix* social engagement and -1 for *Low* social engagement. The *Mix* social engagement appears when a time section contains very similar numbers of *High* (1) and *Low* (-1) datapoints, such that the average on that time section is greater than $-.49$ but lower than $.49$ (as in the equation above). In the ground truth data, this tends to happen at transitions between low and high, but in the prediction data it can also happen when the model is not very stable, the output fluctuating from one social engagement rating to another. These are the three categories for all model’s confusion matrices as seen in Table 4.6 and Table 4.7.

To compute the performance, we compare each rounded window value from the true data to the corresponding time window in the predicted dataset. We evaluate all trained models based on accuracy and F1-score metrics. Accuracy is a measure that shows how often the model’s output is correct. F1-score [Chinchor, 1992] measures how well a model performs, combining precision and recall by their harmonic mean (Equation 4.7). Precision is the number of true positives (true data that is predicted as being true) divided by the number of true positives plus the number of false positives (true data that is predicted as being false); while recall is the number of true positives divided by the number of true positives plus the number of false negatives (true data that

is predicted as being false):

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (4.7)$$

4.5.2 Model configurations

We consider different model configurations for training the model with the derived data (Table 4.3). We compare these configurations to test our assumption that the temporal model (LSTM) and/or pre-training improve the performance of detecting social engagement. We performed these tests using derived input data related to social engagement, as described in Section 4.4.

We considered derived input data because certain behaviours are related to social engagement [Mota and Picard, 2003, Sanghvi et al., 2011]. We pre-processed the raw data to calculate some of these behaviours. We use the derived data to train the machine learning model on four different model configurations (see Section 4.4).

A repeated two-way ANOVA indicated both LSTM and pre-training had a significant effect on the accuracy and F1-score (accuracy - LSTM: $F_{(1,37)} = 58.52, p < 0.001, \eta^2 = 0.613$, pre-training: $F_{(1,37)} = 386.12, p < 0.001, \eta^2 = 0.913$; F1-score - LSTM: $F_{(1,37)} = 14.83, p < 0.001, \eta^2 = 0.286$, pre-training $F_{(1,37)} = 412.74, p < 0.001, \eta^2 = 0.918$), there is also an interaction effect LSTM x pre-training (accuracy $F_{(1,37)} = 11.13, p = .002, \eta^2 = 0.231$, F1-score $F_{(1,37)} = 7.57, p = .009, \eta^2 = 0.170$). This means both LSTM and pre-training have significantly improved the result, and both should be used at the same time to get the best results.

Figure 4.8 A shows the variance of accuracy and F1-score metrics on different model configurations trained on derived data; the numbers on the figure represent the averages of these. As hypothesised, the configuration with both LSTM and pre-training performs the best in terms of accuracy and F1-score average (83.4% and 84.1%). The second best is the configuration where the model is pre-trained but does not have a temporal memory (LSTM). Although its average accuracy and F1-score are considerably higher than the other two configurations, the model results show a high variance compared to the best performing one (PPO+GAIL+LSTM+PreTrain), see Figure 4.8 A. Without considering the outliers, it registers values as low as 9.9% for F1-score and 27.4% for accuracy.

The remaining two models show a low performance: 31.8% accuracy, 40% F1-score for PPO+GAIL+LSTM configuration and 34.2% accuracy, 45.2% F1-score for the PPO+GAIL configuration. This indicates that pre-training has a significant contribution to the model configuration. However, pre-training and LSTM together with PPO and GAIL performs the best across all tested data.

The confusion matrices for all configurations are shown in Table 4.6. The three categories (*Low*, *Mix* and *High*) are a result of the second data post-processing (see Equation 4.6). Unlike PPO+GAIL+LSTM+PreTrain, all other three model configurations (PPO+GAIL+PreTrain, PPO+GAIL+LSTM and PPO+GAIL) have a high values in the *Mix* category: 60, 110 and 102 compared to the actual amount of the *Mix* category: 3. The *Mix* category represents roughly equal amounts of high and low social engagement values (1 and -1). High proportions of *Mix* are therefore likely to indicate a noisy model, the prediction fluctuating from one social engagement rating to another.

	Pred. Act.	Low	Mix	High
	PPO+GAIL+LSTM+PreT	Low	145	2
Mix		2	0	2
High		21	1	97
PPO+GAIL+PreT	Low	136	19	14
	Mix	2	0	1
	High	32	41	47
PPO+GAIL+LSTM	Low	77	65	26
	Mix	2	1	0
	High	60	43	14
PPO+GAIL	Low	44	61	63
	Mix	1	1	1
	High	25	40	54

TABLE 4.6: Confusion matrices for each ML model configuration.

The confusion matrix for each configuration is an averaged confusion matrix from all 38 sessions.

The *Low*, *Mix*, *High* are the categories, denoting high social engagement, mix social engagement and low social engagement. The rows show the actual (*Act.*) data (from the ground truth) and the columns show the predicted (*Pred.*) data (the model’s outcome)

4.5.3 Derived vs raw features

Based on results in deep representation learning [Bengio et al., 2013], we hypothesise that the model trained with raw input data might yield similar results as the models trained with the derived data. The raw features are the base of the derived features. Therefore, an ML model with a complex configuration such as (PPO+GAIL+LSTM+PreTraining), which performed best with derived data, could be able to infer from the raw data and generalise to detect the engagement level in a social interaction [Bengio et al., 2013].

Therefore, we train the best-performing configuration with raw data, following the same procedure to calculate the accuracy and F1-score. The mean values of these metrics are not too low, with 60% accuracy and 63% F1-score, however, there is a very high variance in the model’s predictions (Figure 4.8 B -combined dataset). We collected the data used for training both types of models (with raw and derived data) in two slightly different setups (see Section 4.3.2). Briefly, the first setup (batch 1) has the VC in a different location than in the second setup (batch 2); apart from that, the VC’s gaze behaviour is triggered in the same way in both batches, however, the VC is gazing at different objects in batch 1 compared to batch 2.

We suspected that the VC’s new position (in batch 2) might have influenced the model trained with raw data. This is because the model performs well on the sessions from batch 1 (for both high and low social engagement), but very low on the sessions from batch 2, especially when trying to detect high social engagement. The difference between the two batches is in the VC’s location. Since the input for training the model includes the VC’s location, we consider this a potential reason.

To test this, we separate the results into each of the two batches and into the engagement categories

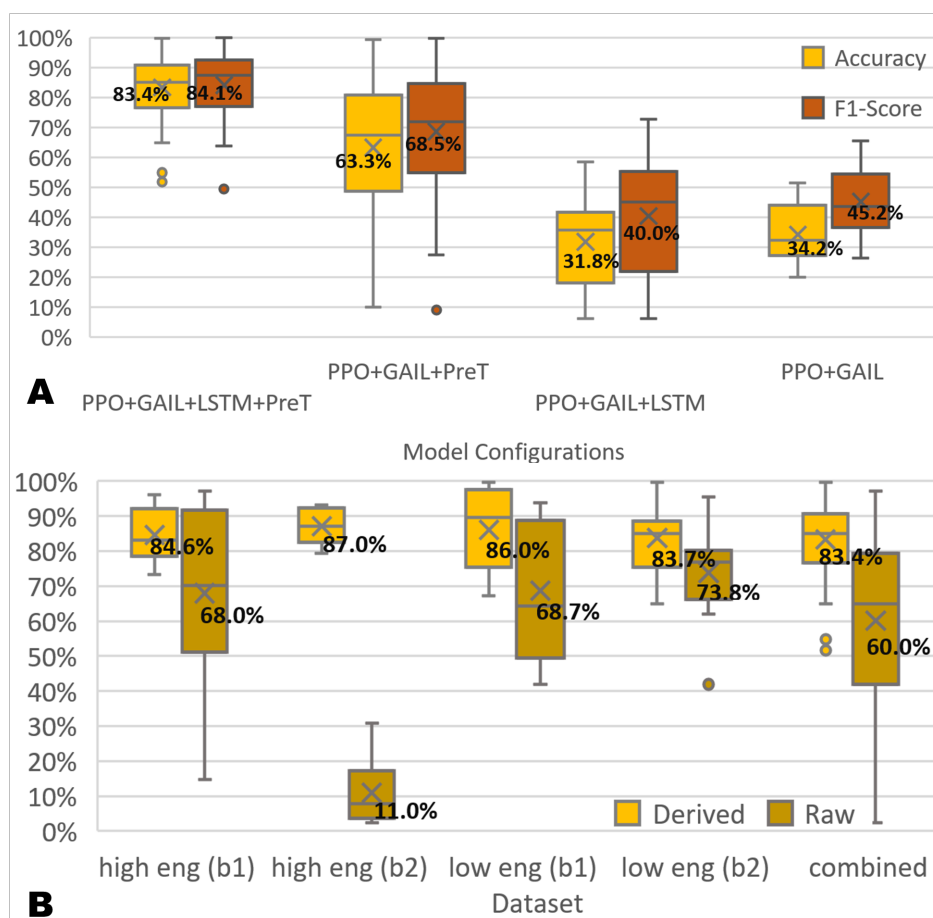


FIGURE 4.8: Accuracy of different ML models configurations using derived and raw data

A: Accuracy and F1 score values for all model configurations trained with derived data on all 38 sessions. *PreT* stands for pre-training. **B:** Accuracy values for PPO+GAIL+LSTM+PreTrain model configurations trained with derived and raw input data on 38 sessions. The *high eng* and *low eng* refer to high, respectively low engagement data based on the human’s annotations. The *(b1)* and *(b2)* represents the first or second batch in which the data was recorded. Finally, *combined* refers to the dataset that puts together all the high, low, and average engagement data. The average engagement data is omitted as there are only 2 sessions, one in each batch. The accuracy for these are: from batch 1, 54.9% and 58.0% for the model with derived respectively raw data; from batch 2, 51.7% and 37.0% for the model with derived respectively raw data.

(low and high). Figure 4.8 B shows a comparison of the two models’ accuracy: one model trained on derived data, and the other on raw data. The F1-score values have a very similar trajectory, hence they are omitted from the figure to not clutter it and placed separately in Figure 4.9. Figure 4.8 B shows raw model’s large accuracy (and F1-score) variance over these 38 sessions. The high engagement data from the second data recording batch register very low accuracy and F1-score values compared to the high engagement data from the first batch. There is no significant difference between the low engagement data from the first and second batches.

We ran a 2×3 Mixed ANOVA analysis (within-group factor *treatment*: raw input data, derived input

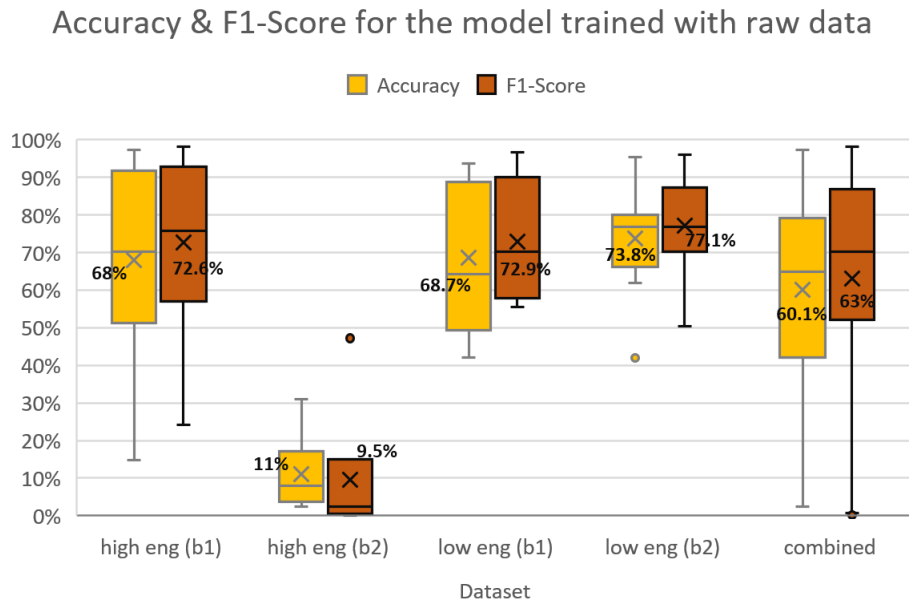


FIGURE 4.9: Accuracy and F1 values for the PPO+GAL+LSTM+PreTrain model configuration trained with raw data

For this we used the raw data from 38 sessions. The *high eng* and *low eng* refer to high, respectively low engagement data based on the human’s annotations. The *(b1)* and *(b2)* represents the first or second batch in which the data was recorded. Finally, *combined* refers to the dataset that puts together all the high, low and average engagement data. The average engagement data is omitted as there are only 2 sessions, one in each batch. The accuracy and F1-score for these are: from batch 1, 58.0% and 62.6% respectively and from batch 2, 37.0% and 35.2% respectively.

data; between-group factor *VR stage*: $S1, S2, S3$). This reveals that the derived data performed significantly better than raw ($p < 0.001$) and that there is a significant *VR stage* effect ($p = 0.045$), but no interaction effect was found ($p = 0.172$). Post-hoc Tukey test reveals that the model performed significantly better for VR stage $S3$, as compared to VR stage $S2$ ($p = 0.035$). No other effects were found between the VR stages.

Table 4.7 contains the confusion matrices for models trained on derived and raw data split based on the data collection batch. The model trained on raw data fails to detect a large proportion of the *High* social engagement parts, mostly miss-predicting them as low social engagement. This model also shows a much higher fluctuation of social engagement rating per 1-second window. This is illustrated in the high amount of predictions for the *Mix* social engagement, 40 (20 + 1 + 19 in Batch 1) and 45 (26 + 1 + 18 in Batch 2) compared to the actual value of 4 (2 + 1 + 1 in Batch 1) and 3 (2 + 1 + 0 in Batch 2).

The model trained with raw data might have learned very specific features, for example, the exact position of the VC. If that condition is not fulfilled (the VC is not positioned in the same location or has a changing position), then the raw data model incorrectly predicts the engagement level. This is a problem as it is very common in games to have VCs that would move in the environment.

There could be a possible solution to improve the raw model’s performance while keeping the VC active in the scene. To do this, more data needs to be collected with the VC in different locations

		Derived Data			Raw Data		
		Low	Mix	High	Low	Mix	High
Pred.	Act.						
	Batch 1	Low	137	2	15	112	20
Mix		2	0	2	2	1	1
High		26	1	103	30	19	82
Batch 2	Low	152	3	27	148	26	8
	Mix	1	0	2	2	1	0
	High	16	1	90	82	18	6

TABLE 4.7: Confusion matrices for PPO+GAIL+LSTM+PreTrain model configuration

The confusion matrix for the models trained with derived and raw data. The confusion matrix for each model is an averaged confusion matrix separated into two data collection batches: Batch 1 (with 18 sessions) and Batch 2 (with 20 sessions).

The *Low*, *Mix*, *High* are the categories, denoting *High*, *Mix* and *Low* social engagement. The rows show the actual (*Act.*) data (the ground truth) and the columns show the predicted (*Pred.*) data (the model’s outcome)

and using more participants to interact with the VC. This might decrease the variance in accuracy for the raw model. However, the process of recording the data and training the model is very expensive and time-consuming, making it unfeasible for a game production process. The use of psychologically inspired derived features is therefore a better approach within the practical time and budget constraints of game development.

4.6 Generalisation: ML pipeline for social attitude detection

In this section we go over the pipeline used to detect social engagement (Section 4.3), generalising it to be used for detecting other social attitudes detection (such as sympathy, affection or aggression).

Both the data collection and the data annotation take place in VR. First, a user interacts with the VC and their behavioural data is collected (see Figure 4.2 A). Next (Figure 4.2 B), a human annotator labels the presence of a social attitude while watching a playback of the user-VC interaction in VR (see Section 4.3).

The data from the user-VC interaction is the base data for training the ML model. It consists of data about the user’s and VC’s activity, such as the movement (head and hands position and rotation), interaction with other objects or with each other, and so on. By performing this data collection in VR, we are able to create a situation that is as close as possible to real gameplays and also to real social interactions. The data collected in this step can be different to the one we collected for social engagement detection (Section 4.3); it should contain relevant data for the specific social attitude (e.g. eye or pulse information).

This data is then played back in VR to be labelled for training the ML model. Because of this, it should contain instances of the social attitude’s presence (positive value) and its absence (negative value). Thus, the VR scenario needs to contain situations that allow both positive and negative examples of a participant’s social attitude.

The human annotator is a key figure in this pipeline. They have the ability to look at the behaviours from the interaction and choose the ones that resemble the complex social attitude the ML model will learn to detect. In game development, a creative director could be the annotator. They use their artistic vision on the final product and decide what social attitude is important to be detected in a particular scene in the game, while a user is socially interacting with a VC/NPC.

To decide this, instead of having to provide a concrete definition of the social attitude, the annotator labels it while observing the playback interaction. In our social engagement example, they can perform this action by pressing a ‘plus’ button on the VR controller when they see the attitude (social engagement) and a ‘minus’ button when there is a lack of it (Figure 4.7 B). This way, the annotations conceptualise the complex and abstract activity (social attitude). Then, the trained ML model will detect this activity during social interaction. Other features could be included in the labelling task in VR to ease and improve the outcome based on the social attitude (e.g. video feed overlay).

By annotating in VR, the annotator is able to make full use of their social cognition skills as they would do in a real-world social interaction. By performing it in a game environment, it can become a game design task, in which a designer can judge the interaction as it would fit into the real gameplay.

The ML model trains on the dataset: the user’s data and the annotator’s labels as ground truth data (Figure 4.2 C). The ML model replicates the human annotations by using an imitation learning algorithm approach, thus mimicking the human intuition of marking an attitude within a social interaction (Figure 4.2 D).

After training the model for detecting, for example, aggression, it can be applied to different scenarios. The model outputs whether aggression is present or absent based on the input data from the interaction (Figure 4.2 E, F & G). Finally, the output can be manipulated and used in real-time in applications (e.g. games) to trigger various actions or behaviours based on the designer’s vision. For instance, when the player is detected to be too aggressive, the NPC could stop talking and could be animated to reflect the behaviour received; when the player is showing empathy, the NPC will start talking again and their animation would change indicating that.

4.7 Limitations and discussion

The exact results of the social engagement detection presented here could be difficult to replicate without the same annotator. However, the aim of the project was not to create a general detection model for social engagement (or other social attitudes) because individuals often have their own standards of what counts as engaged or not [Glas and Pelachaud, 2015]. In our case study of detecting social engagement using the proposed pipeline, instead of explicitly defining the present or absent criteria of a certain social attitude, we rely on the annotator’s ability to label it. In game companies, this annotator role should be taken by their creative director, making the labelling itself part of the creative process. In other words, we aim to detect the High/Low social engagements that are modelled based on the creative game designer’s (the annotator’s) markings. During the

VR interaction playback, they would be labelling the behaviours identified in players which are related to the presence or absence of social engagement. Thus, when real-life players exhibited those behaviours during gameplay, certain events (NPC behaviours, or change of game environment) could be then triggered.

For this work, the annotator marked the data in a binary way: either high or low social engagement data. This can be a limitation of our approach since social attitudes are not necessarily binary. We decided to go with this approach because the algorithms within the Unity ML agents have the requirement of binary input data hence we kept the annotations and the model's predictions binary. However, the model predicts at a higher frequency (9 to 10 frames per second) compared to other work such as Yu et al.'s study [Yu et al., 2004], where predictions take place per each utterance, or Bohus and Horvitz's work [Bohus and Horvitz, 2014], where they use a 5- seconds window to forecast disengagement. In this case, the game system can use the prediction model at a finer level. For instance, the game designer can choose to calculate an average social engagement over time spans of 5 seconds (as in Bohus and Horvitz's study [Bohus and Horvitz, 2014]). For that time window, there will be a total of 45-50 (5x9, 5x10) predictions which can be used to calculate a fine-grained outcome, rather than a binary one.

We ensured the annotator was making judgements only on the data available to the ML model. Thus, the annotator labelled the data in VR, they could have access only to the data available to the ML model. If the annotator was making judgements using data inaccessible to the ML algorithm, we argue that the algorithm cannot learn from it, as described in Gillies et al. [Gillies et al., 2015]. However, we based this decision on prior literature and we did not attempt to annotate the interactions on more data compared to the one used for training the models. For future work, we could rerun the annotation process on the user-VC interactions giving the annotator more data (i.e., the audio feed).

We trained the ML models using data from all those participating in the interaction (the user and the VC). We took this decision based on prior literature (see Chapter 2.2.3): as the VC should be able to assess the social engagement based on the interaction with the user. Additionally, prior work shows that ML models that use only one person's data have lower performance than the ones taking into account both people's data [Dermouche and Pelachaud, 2019a]. However, a limitation of our work in this chapter is the comparison of the ML models trained on data from the user-VC interaction versus ML models trained solely on the user's data. Based on prior work, we believe that including data from the VC improved our ML model performance and made it more ecologically valid. Nevertheless, this was not tested with our dataset, and it is open for future work.

We collected data from participants in a western city who volunteered to take part in the study hence they might have an interest in XR. For this reason, the behaviour and social attitude expression recorded are linked to the cultural background. A potential further work could be to run studies with participants from other backgrounds to enrich and compare the dataset and the detection model. Even though we collected data from 38 sessions and from 13 participants, the dataset was not very large. We also selected features that were readily available to train the ML model. It would

be interesting further work to consider a different array of features available from more cutting-edge hardware.

We run the data collection in two batches to investigate the effect of the agents' gaze at various key objects. This aspect is out of scope for this chapter, however, recording the data in two batches with a distinct VC location for each allowed for a more diverse dataset (see Section 4.3.2). As future work, we could diversify it even more by assigning a different VC starting point for each participant.

Despite our limitations, we received very positive comments from our industry collaborators. The industry collaborators helped to create the tool and as a retrospective note, they commented on how non-verbal communication is at the centre of the tool, and that the use of nonverbal behaviour widens the applicability of this work for other types of games and applications within the entertainment and games industries. During the collaboration we chose the social engagement case study, however, one of the game companies applied the pipeline for developing a VR karaoke-style application where the user would sing along with an NPC singer, which will change their attitude depending on the social attitude of the player in real-time. Likewise, they didn't define the social attitude but the annotation was based on the way the players sang, moved, performed, and how involved they were in the experience. The CEO of the game company commented on their experience of building the VR karaoke app using the immersive ML pipeline: *'The ML project was very interesting to be a part of, seeing it grow from a very simple idea into something quite sophisticated. What impressed me the most was seeing the same principles used in a 19th century narrative game also grimly applied to a modern karaoke game. This generalisation convinced me of the merit of the approach taken. From experience, I think it is relatively straightforward to get a system working on one context, but to reapply the same principles in a fundamentally different context proves its true worth.'*

4.8 Conclusion

In this chapter, we present our collaborative work with two game companies (Dream Reality Interactive and Maze Theory) to develop a pipeline with immersive data collection and annotation in VR for training an ML model. We design the pipeline to support the games industry's creative design process and to be integrated into production-ready VR games for the consumer market.

The pipeline is used to train an ML model to detect social attitudes, such as sympathy, social engagement, or aggression, using a reinforcement learning (PPO) approach with rewards based on an imitation learning algorithm (GAIL). In this study, we show how the pipeline is used to train an ML to detect social engagement.

We consider different model configurations and input data for training the model: derived data and raw data. The model using derived data performs the best, while the model based on raw data is not able to generalise to different VC positions. The model configuration that yields the highest accuracy and F1-score (83.4%, 84.1%) is based on a reinforcement learning algorithm (PPO) with imitation learning rewards (GAIL) implementing a temporal memory (through LSMT) and a

pre-training algorithm. The other model configurations perform poorly, the outcome being a rapid change between high and low social engagement values in a short period of time.

The proposed work contributes to the field of socially responsive VCs, offering a design-by-example tool for immersive ML, to detect abstract social attitudes in VR social interactions. This could be useful in designing social interactions in VR games or in other immersive experiences (simulations, training, social platforms), where the user can interact with the VC using their own bodies, as they do in everyday life. This opens opportunities for novel input interactions, game mechanics or VC's behavioural models that are related to the rapport/empathy between the user(s) and the VC.

In this chapter, the focus is on high-level behaviours in social interaction, detecting social engagement with a case study in the games industry. Next, we consider a different industry field that relies extensively on virtual environments (2D displayed and immersive virtual environments): the remote-working field. In the next chapter, we focus more on avatars rather than on virtual characters as they are more frequently encountered in the work industry and remote meetings. We are interested in the impact of different appearance styles of avatars, comparing personalised cartoon and realistic avatars in real-work meetings in mixed reality. Even though our application is specific to a certain industry (remote working) the results from this study are also applicable to virtual characters in other immersive virtual environments.

5

AVATARS: The Appearance Impact in Real-World Meetings

In this chapter¹ we present a within-subjects study that examines the effects of realistic and cartoon avatars on communication, task satisfaction, sense of presence, emotional state perception, and useful cues 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 or cartoon face. Half of the groups started with the realistic condition and half with the cartoon condition; all groups switched conditions halfway through the study. Results showed that participants using realistic avatars first may have had higher expectations and more errors in perceiving their colleagues' emotional states. Participants using cartoon avatars first 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' emotions, regardless of avatar style. When starting with realistic avatars, participants rated gaze as more useful than facial expressions, while when starting with cartoon avatars, both gaze and facial expressions were rated as the least useful. Results also suggested that participants had more errors when perceiving negative emotional states in their colleagues, with this trend appearing for most emotional states but depending on the avatar style order. Implications of these findings for mixed and virtual reality meetings are discussed. This work contributes to the field of remote collaboration by providing

¹Work published in: *Dobre, Georgiana Cristina, Marta Wilczkowiak, Marco Gillies, Xueni Pan, and Sean Rintel. "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. 2022 <https://doi.org/10.1145/3491101.3519628>*

insights from longitudinal data on the impact of avatar appearance on various aspects of work meetings in virtual environments.

5.1 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 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 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, and facial expressions) and the avatar's actual behaviour. Cartoon styling, 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 2.3).

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

In summary, most of what we know about avatar appearance in meeting-style 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.

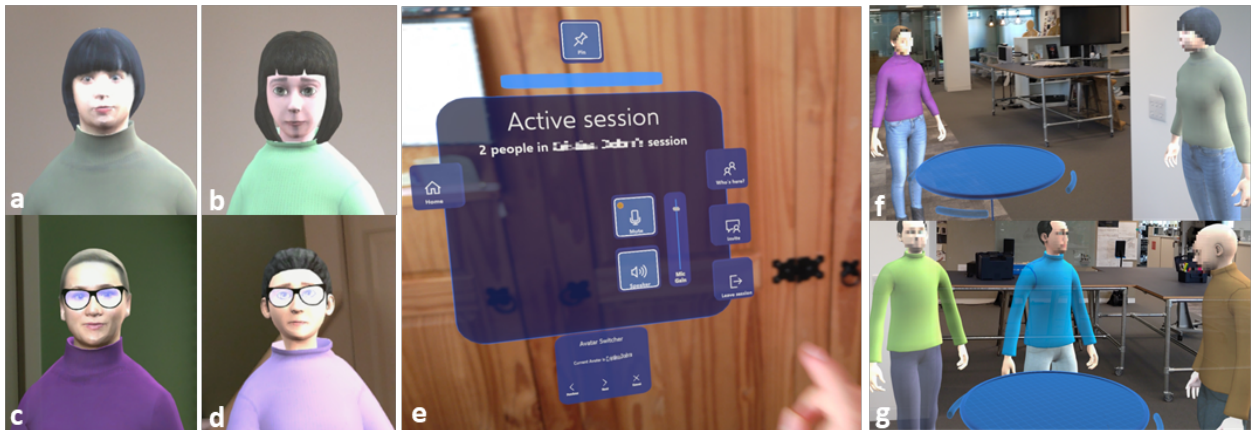


FIGURE 5.1: The Avatars and the Mixed Reality environment

Example of realistic (a,c) and cartoon (b,d) avatar upper bodies. The main menu controls the session (e). A meeting between two participants in realistic avatars (f) and three in cartoon avatars (g) with the adjustable blue table marking the centre. Participants gave consent to use their avatars publicly.

In this chapter, we address this gap by investigating how people feel about using avatars with different appearance styles in immersive meetings over an extended period of time. For two to three weeks, six groups of co-workers (14 people) from Microsoft Corporation 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 the novelty factor waned. Specifically, we were interested in examining the functional communicative value, task satisfaction, presence, and the self-reported and perceived emotional states of individuals during immersive virtual meetings.

In the following sections, we introduce the research questions in Section 5.2 and detail our methodology in Section 5.3. We then present our data analysis and results in Sections 5.4 and 5.5, respectively, before embarking on a discussion of these findings in Section 5.6 where we also report the limitations. Finally, we conclude in Section 5.7.

5.2 Research Questions

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 (see Section 2.3). 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., 2020, 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.

In this chapter, we cover seven research questions (RQs) split into two sets. The first set of questions covers the effect of the avatar style on communicative value, task satisfaction, and presence. The remaining set of research questions covers the ability to recognise others' emotional states. For ease of expression in the chapter, we cover the two RQ sets separately in the Results (Section 5.5), but bring them together in the Discussion (Section 5.6).

The research questions are as follows:

SET ONE

How do the avatar representations influences:

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.

SET TWO

RQ4: Does the avatar representation change the self-reported emotional states overall and over time?

RQ5: Does the type of avatar style affect how accurately people perceive the co-workers' emotional states (a) overall and (b) do they improve overtime?

RQ6: Do positive or negative emotional states affect how accurately they are perceived by others?

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

5.3 Methodology

Device and application. The study run in Mixed Reality using the HoloLens 2 (HL2) device (microsoft.com/en-gb/hololens). We built a networked application using Unity3D game engine (unity.com) where users see a hologram of a blue table (Figures 5.1f and 5.1g) and a control menu

(Figure 5.1e). The table is adjustable and represents the centre of the meeting, all other participants in the meeting are located in space around the table. The control menu allows the participants to go to the ‘Home’ menu and to create a new meeting, see who is in the current meeting, join a meeting, mute themselves, adjust their microphone gain, switched their avatar, leave the meeting and quit the application.

Avatars. Participants used full-body avatars in both a cartoon and a realistic 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 (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 (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 male and two female, 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 participants had different coloured clothing (as shown in Figures 5.1c-e).

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. The heads’ facial animation was generated using a simple lip-flapping script based on voice amplitude, as well as a blinking animation. However, due to time constraints, the avatars did not have a seating animation or seated static position, so participants were instructed to stand for the duration of their meetings.

Participants. 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). We offered a charity donation of £75.00 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 5.1). 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

#	Style W1	Style W2	Gender	Size	Sessions	Q.
1	Realistic	Cartoon	F, F	2	10	20
2	Realistic	Cartoon	F, M	2	10	18*
3	Realistic	Cartoon	F, F, Non-Binary	3	8	24
4	Cartoon	Realistic	M, M, M	3	10	30
5	Cartoon	Realistic	M, M	2	6	12
6	Cartoon	Realistic	F, F	2	10	20
Total	3-R;3-C	3-C;3-R	7-F; 6-M; 1-Non-B	4-Dyads; 2-Triads	54	124

TABLE 5.1: Details on the participants and the data collected

Details on the group size, participant’s demographic, avatar order and sessions. There are 18(*) (instead of 20) questionnaires filled in for group 2 because, due to a technical error, there is a missing set of questionnaires from the last session using the cartoon avatars.

10 – 15 minutes. These meetings often took the form of daily stand-ups, status reports, or daily team catch-ups.

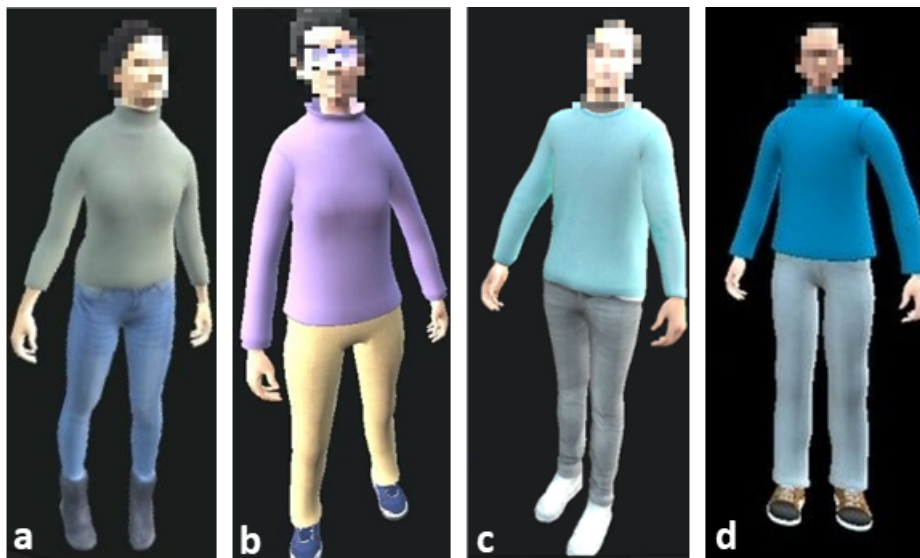


FIGURE 5.2: Personalised full-body avatars

Example of some of avatars used (a-d); from left to right: female realistic(a), female cartoon(b), male realistic(c), male cartoon(d). The faces are blurred as not all participants provided consent for using their personalised avatars publicly. This figure shows examples of the full-body avatars and the difference in shirt colour for each avatar.

Dataset.

We collected data from questionnaires, daily meetings, and one focus group from each group. The participants completed consent forms and the following questionnaires: demographic, onboarding (covering their expectations of having meetings in MR), and a daily questionnaire that they completed after each meeting. In the daily meetings we collected head and hand movements, audio amplitude, and proportion of speaking participation, but to maintain confidentiality we did not capture actual speech data (audio or transcribed).

In Table 5.1 we structure the details regarding the participants and the groups they were part of. 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 (total number of *Sessions* in Table 5.1), 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. Half of the groups used cartoon avatars first, while the other half used realistic avatars first. 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.

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 address RQs 1 – 3. For all questions see Table 5.2.

The second part of the questionnaire focused on the recognition of emotional states and the usefulness of cues for perceiving these emotional states. Four emotional states were selected based on the UWIST mood checklist [[Matthews et al., 1990](#)]: optimistic, focused, annoyed, and stressed. We selected these emotional states as they were relevant to the study setup (workplace meetings). Also, we did not choose a larger subset of emotional states to keep the length of the questionnaires short. Participants were asked to fill these questionnaires every day after the meetings and by using a longer questionnaire, we were expecting to see a drop in completing them or in dropping participants from the study [[Galesic, 2006](#)]. Thus, participants were asked to rate their own emotional states and the perceived emotional states of their colleagues on a 1 – 7 Likert scale ranging from "strongly disagree" to "strongly agree" for each emotional state. Participants in triads rated the perceived emotional states of the other two participants. Next, they were asked to rank 5 cues in order of their usefulness for recognising emotional states 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.

Procedure.

After providing their consent, participants completed the demographic and onboarding 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.

The procedure for each session was as follows: the participants opened the 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 (the rest of the group appeared around this table). One group member created a meeting and added the others. 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. Following the meeting, the researcher reminded them to complete the questionnaires for that session. This process was repeated until the final session.

5.4 Data Analysis

5.4.1 Reverse Coding

After each meeting session, participants filled out two questionnaires, one on communication and one on emotional states. For the emotional states questionnaire, participants had to rate their emotional state on a scale from 1: *Strongly Disagree* to 7: *Strongly Agree*. They rated four emotional states: *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 emotional states, we utilised a reverse-coding technique for the negative emotional states (*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 emotional state *Stressed* would be transformed into a rating of 3 ($8 - 5$). These reverse-coded ratings were then utilised in our data analysis to answer *RQ4*, as detailed in Section 5.5.2.

5.4.2 Accuracy of perceived emotional states

We calculated the accuracy of the perceived emotional state by computing the error that participants had when perceiving the emotional states of their colleagues. We determined the error by mapping the absolute value of the difference between the self-reported emotional state and the perceived rating of the emotional state 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 emotional state, 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 0.667.

We calculated this error for each pair of participants in a group. In dyads, we considered the error for each participant in perceiving the emotional state of their colleague: P1's error in perceiving P2's emotional state (*P1_to_P2*) and P2's error in perceiving P1's emotional state (*P2_to_P1*). In triads, we considered each participant's error in perceiving the emotional states 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 5.5.2). We also used it to address RQ5b on participants’ ability to accurately perceive their colleagues’ emotional states over time (Section 5.5.2). Finally, we used it to test RQ6 and the relationship between negative self-reported emotions and higher errors in perceived emotional states (Section 5.5.2).

5.4.3 Avatar Order

We took into account the order in which the avatar styles were used since this might influence the participant’s behaviour [Ma and Pan, 2022]. Half of the participants used the Cartoon avatar in Week 1 (W1) and then used the Realistic avatar in Week 2 (W2). The other half used the Realistic avatar in W1 and then used the Cartoon avatar in W2. To avoid confusion about the starting avatar style and the avatar style used, we use the following naming convention for each condition. 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).

5.5 Results

5.5.1 Part One: Communication, Tasks and Presence

#	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 computerized 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

TABLE 5.2: The items in the daily questionnaire

Participants answered this questionnaire on a 1 – 7 Likert scale. The star(*) items showed significance. *RQ* stands for *Research Question*. *m* and *sd* stand for mean and standard deviation, showing the descriptive statistics for each question.

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.

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 Repeated Measures ANOVA to assess the effect of realism on the data. The descriptive statistics for this analysis can be found in Table 5.2, and a boxplot representation of the results for each question can be found in Figure 5.3A.

RQ1 Functional communicative value. On average, for realistic avatars, 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 ($F(1, 13) = 7.14, p = .019, \eta^2 = .355$) and Q12 ($F(1, 13) = 5.5, p = .036, \eta^2 = .296$), but not for Q8 ($F(1, 13) = 3.53, p = .08, \eta^2 = .217$) or Q3 ($F(1, 13) = .718, p = .41, \eta^2 = .52$), Figure 5.3A RQ1.

There was a significant interaction effect between avatar style and order of use on participants' ratings of nonverbal behaviour appropriateness in Q11 ($F = 13.01, p = .004, \eta^2 = .52$). 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 Figure 5.3A, as well as in RQ11 and RQ12. 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

To account for the temporal aspect, we calculated the regression statistics for each dependent variable, with respect to which avatar type the participants embodied. To do this, we computed the data for each avatar type, combining W1 and W2 (shown in Figure 5.3B). We then present the data based on the avatar usage order (either Cartoon—Realistic (CR) or Realistic—Cartoon (RC)) in Figure 5.3C and 5.3D.

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, but

not the Realistic avatars (shown in Figure 5.3B 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 Figure 5.3B for RQ2). However, when considering the order, we see a significant decrease in participants' responses for the Cartoon avatars in the Cartoon—Realistic order ($R^2 = .13$, $F(1, 29) = 4.18$, $p = .05$, shown in Figure 5.3C 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, it was found that the appearance of the avatar mattered less over time for participants using Cartoon avatars ($R^2 = .1$, $F(1, 59) = 6.67$, $p = .01$, Figure 5.3B Q6). Additionally, participants using Realistic avatars reported fewer obvious unnatural nonverbal behaviours over time ($R^2 = .1$, $F(1, 61) = 6.22$, $p = .01$, Figure 5.3B Q10). No other significant findings were discovered when examining data from both weeks (W1 and W2).

When examining the order in which avatar 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: $R^2 = .27$, $F(1, 29) = 11.06$, $p = .002$, Figure 5.3C Q6). When Cartoon avatars were used in the second week (Realistic—Cartoon order), there was a decrease but it was not significant (W2: $R^2 = .03$, $F(1, 28) = .74$, $p = .39$, Figure 5.3D Q6). For the Cartoon—Realistic order group, there was also a significant drop for Realistic avatars in the second week (W2: $R^2 = .23$, $F(1, 29) = 9.1$, $p = .005$, Figure 5.3C Q6), with the opposite trend observed for the Realistic—Cartoon order group in the first week, but it was not significant (W1: $R^2 = .04$, $F(1, 30) = 1.38$, $p = .24$ (Fig 5.3D 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: $R^2 = .25$, $F(1, 30) = 10.06$, $p = .003$, Figure 5.3D, Q10), but not during the Cartoon—Realistic order group (W2 $R^2 = .07$, $F(1, 29) = 2.09$, $p = .16$, Figure 5.3C, Q10). No other significant findings were discovered for RQ3 on the other questions.

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). Co-workers used Cartoon and Realistic avatars for 2 – 3 weeks and based on their questionnaire responses, we found important outcomes. 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. Over time, participants reported an improvement in identifying their colleagues while embodying Cartoon avatars (Q8 Figure 5.3B). When participants used Cartoon avatars first, they reported that the avatar's appearance mattered less to them over time. This trend appeared

for Cartoon and Realistic avatars during the Order Cartoon—Realistic, but not for the Order Realistic—Cartoon (Q6 Figure 5.3C). At the same time, when participants embodied first the Cartoon style avatars, they reported that their appearance affected less their comfort over time (Q5 Figure 5.3C). When participants used the Realistic avatars for the first time, they reported less unnatural nonverbal behaviours over time (Q10 Figure 5.3D); this trend is not significant for Order Cartoon—Realistic, not for when participants embodied Cartoon avatars.

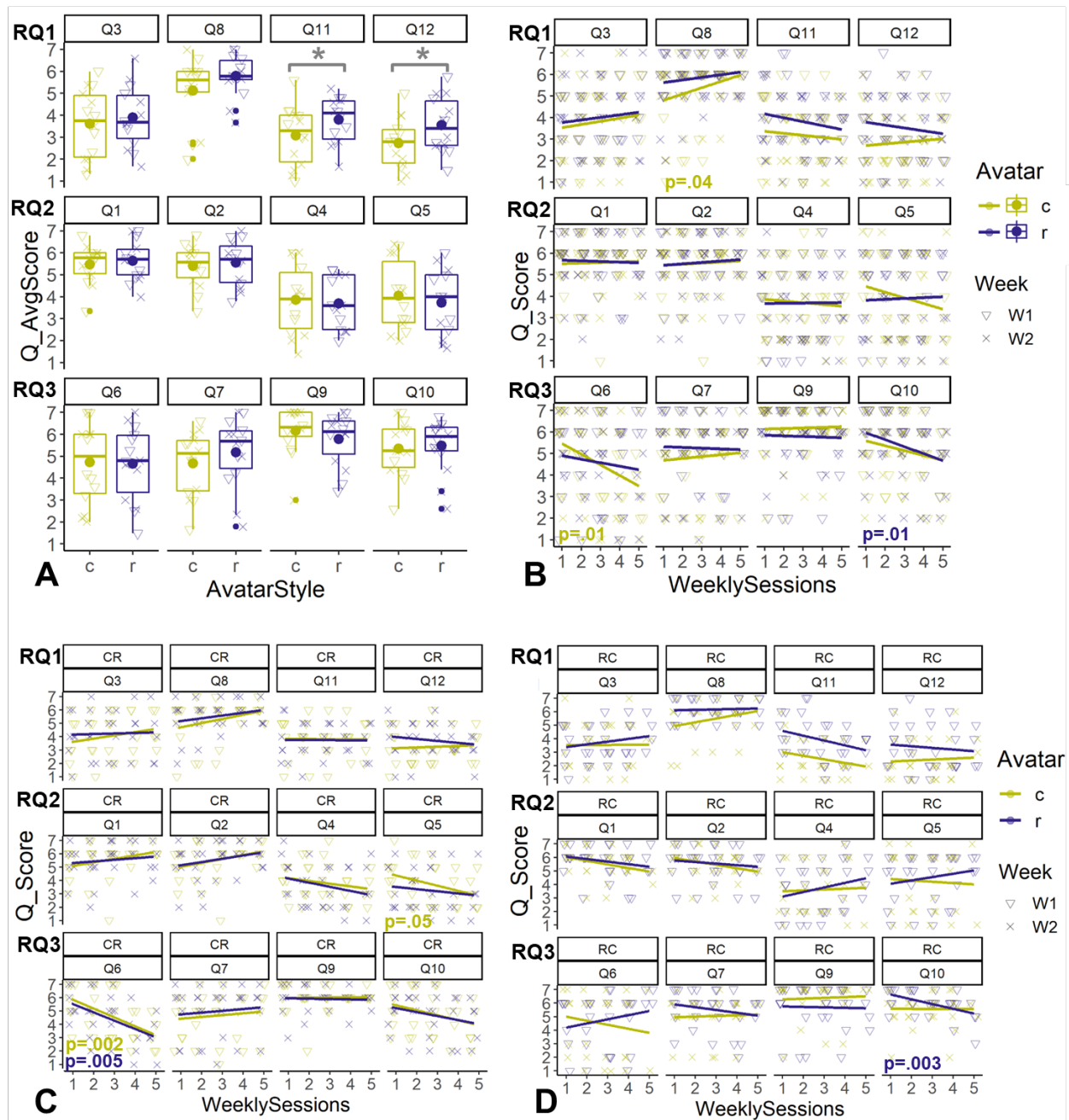


FIGURE 5.3: Responses to the questionnaire from Table 5.2 overall and over time

Panel A: Boxplots for each question separated by the avatar style; **Panel B-D:** Scatter plots showing each question score over time separated by the avatar style. The X-axis represents the weekly sessions in chronological order. In **Panel C-D**, the question score is separated by the order (Cartoon—Realistic and Realistic—Cartoon).

5.5.2 Part Two: Perceived and Self-Reported Emotional State

RQ4: Self-Reported Emotions Overall and Over Time

Participants self-reported their emotional states daily after each meeting. We are interested to investigate the self-reported emotional state while participants embodied different avatar styles and how this changes over time.

To compare the effects of Cartoon and Realistic avatars on emotional states, we first reversed the ratings for Annoyed and Stressed for all participants (as explained in Section 5.4). We then 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 Emotional State, and no significant difference was found ($p = 0.98$; Cartoon: mean = 4.27, variance = 1.13; Realistic: mean = 4.26, variance = 1.12).

Previous research has suggested that the order in which participants experience an avatar can influence their subjective experience, so we also conducted separate paired t-tests between Realistic and Cartoon avatars for each order. Using the same data where the self-reported ratings were reversed for Annoyed and Stressed, we found no significant difference in the self-reported emotional state for the Cartoon—Realistic order ($t(27) = .45$, $p = .65$ Cartoon W1: mean = 4.75, variance = 2.07; Realistic W2: mean = 4.69, variance = 1.83). However, for the Realistic—Cartoon order, participants reported more positive emotions when using Realistic avatars in Week 1 compared to Cartoon avatars in Week 2 ($t(27) = 2.81$, $p = .009$, Realistic W1: mean = 4.9, variance = 1.29; Cartoon W2: mean = 4.23, variance = 1.81; see Figure 5.4A). We also conducted paired t-tests for each Emotional State and for each Order, comparing Cartoon and Realistic avatars, but no significant results were found (see Appendix D for detailed statistics).

Next, for each participant, we calculated the average self-reported score for each avatar and each Emotional State. Because we took into account the Emotional State, we did not reverse the ratings for Annoyed and Stressed. A paired t-test was conducted to compare the self-reported Emotional State ratings for Cartoon and Realistic avatars, regardless of Order. The results showed that participants self-reported feeling more Optimistic in meetings using Realistic avatars compared to Cartoon avatars ($t(13) = 2.53$, $p = .025$; Cartoon: mean = 4.6, variance = 1.3; Realistic: mean = 5.06, variance = 1.68; see Figure 5.4A). There was no significant difference for the other emotions (Focused $t(13) = -.11$, $p = .92$, Annoyed $t(13) = -.57$, $p = .58$, or Stressed $t(13) = -.37$, $p = .71$).

We also analysed the participants' self-reported emotional states over time while using either Cartoon or Realistic avatars. The data was split by avatar style and Weekly Session of avatar use, and regression analyses were conducted on the self-reported emotional states over time for Order Cartoon—Realistic (CW1, RW2) and Order Realistic—Cartoon (RW1, CW2). A significant result was only found for Realistic avatars used in the first week (Order Realistic—Cartoon). Specifically, while using Realistic avatars for the first time, participants self-reported feeling less Optimistic over time ($R^2 = .14$, $F(1, 30) = 4.93$, $p = .034$, Figure 5.4C) and more Stressed over time ($R^2 = .16$, $F(1, 30) = 5.66$, $p = .024$, Figure 5.4C). There were no significant results for the Focused and Annoyed emotional states or for the Cartoon avatars (see Appendix D for detailed statistics).

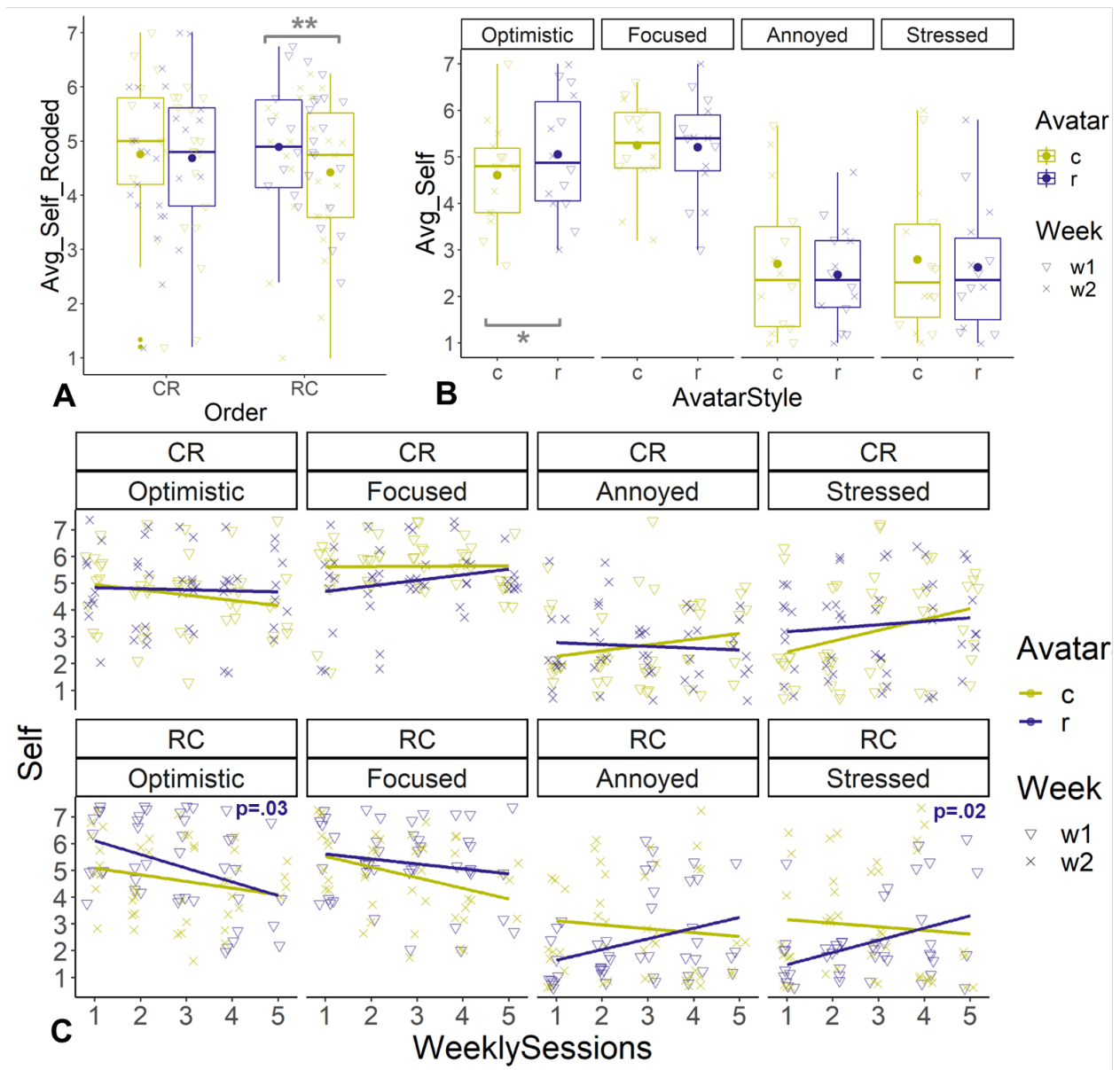


FIGURE 5.4: Self-reported emotional state ratings, overall, over time, by emotional state and by Order

Panel A: Box plots showing the average self-reported emotional state ratings, with reverse coding for Stressed and Annoyed. The data is split by the order of Cartoon—Realistic and Realistic—Cartoon, and by the avatar style. On the Y-axis, the range is from 1 (representing Strongly Disagreeing) to 7 (representing Strongly Agreeing) of having a certain emotional state. The X-axis shows the order (Cartoon—Realistic and Realistic—Cartoon). **Panel B:** Box plots showing the average self-reported ratings from each participant by emotional state and avatar style. On the Y-axis, the average self-reported ratings range from 1 (representing Strongly Disagreeing) to 7 (representing Strongly Agreeing) of having a certain emotional state. The self-reported ratings are separated by the avatar style (Cartoon or Realistic) that participants were embodying, as shown in the box plots. **Panel C:** Scatter plots showing the self-reported ratings from each participant by order (Cartoon—Realistic in the top row and Realistic—Cartoon in the bottom row) and emotional state. The avatar style is further represented by a fitted line. On the Y-axis, the self-reported ratings are shown, and on the X-axis, the weekly sessions are shown chronologically from 1 to 5.

RQ5a: Accuracy of perceived emotional states overall

Next, we were interested in the effect of avatar type on the accuracy of people's perception of their co-workers' emotional states.

We conducted a repeated measure 2×4 ANOVA on the Mapped Error using avatar style (Realistic, Cartoon) and emotional state (Optimistic, Focused, Annoyed, Stressed) as within-subjects factors, and order (Cartoon—Realistic and Realistic—Cartoon) as between-subjects factor. Our results showed a significant difference in the error of perceived emotional states when using Cartoon versus Realistic avatars ($F(1, 18) = 5.13$, $p = .036$, $\eta^2 = .22$). Specifically, participants perceived their colleagues' emotional states with fewer errors when using Cartoon avatars ($M = .222$) compared to Realistic avatars ($M = .248$).

We found that there was a significant interaction effect between the order in which the avatars were used and the type of avatar ($F(1, 18) = 5.91$, $p = .026$, $\eta^2 = .247$). This is evident in the higher error rate for Realistic avatars when they are used in the Realistic—Cartoon order compared to the Cartoon—Realistic order. Specifically, when participants used Realistic avatars in the first week, their mean error rate was .270, while it was .226 when Realistic avatars were used in the second week (see Figure 5.5A).

There was also an interaction effect between the order, the style of the avatar, and the emotional state of the participant ($F(3, 18) = 3.71$, $p = .017$, $\eta^2 = .171$), as shown in Figure 5.5B. When examining the data by emotional state, we found that the error rates varied depending on the order and style of avatar used. For Cartoon avatars, the error rate was generally higher for the positive emotional states (Optimistic and Focused) in the Realistic—Cartoon order compared to the Cartoon—Realistic order. However, for the negative emotional states (Annoyed and Stressed), the pattern was reversed, with the error rate being higher for the Cartoon—Realistic order. For Realistic avatars, the error rate was higher for all emotional states except for the Optimistic state.

To investigate the interaction effect between avatar style and order, we conducted a post-hoc analysis using a paired t-test to compare errors made by participants in the Cartoon—Realistic and Realistic—Cartoon orders. The results showed that there was no significant difference in the Cartoon—Realistic order ($t(9) = .11$, $p = .91$; Cartoon W1: mean = .228, variance = .006; Realistic W2: mean = .226, variance = .003). However, in the Realistic—Cartoon order, participants made more errors in perceiving their colleagues' emotional states while using the Realistic avatar in the first week and then the Cartoon avatar in the second week ($t(9) = 3.79$, $p = .004$; Realistic W1: mean = .27, variance = .004; Cartoon W2: mean = .21, variance = .001).

We also conducted a two-factor ANOVA with avatar style (Cartoon and Realistic) as the dependent variable and order as the between-subjects factor to further explore the interaction effect between avatar, order, and emotional state. The results revealed significant differences between Cartoon and Realistic avatars for the Annoyed ($F = 4.41$, $p = .05$) and Stressed ($F = 5.32$, $p = .033$) emotional states, but no significant differences were found for the Optimistic ($F = 2.77$, $p = .11$) or Focused ($F = .11$, $p = .74$) emotional states.

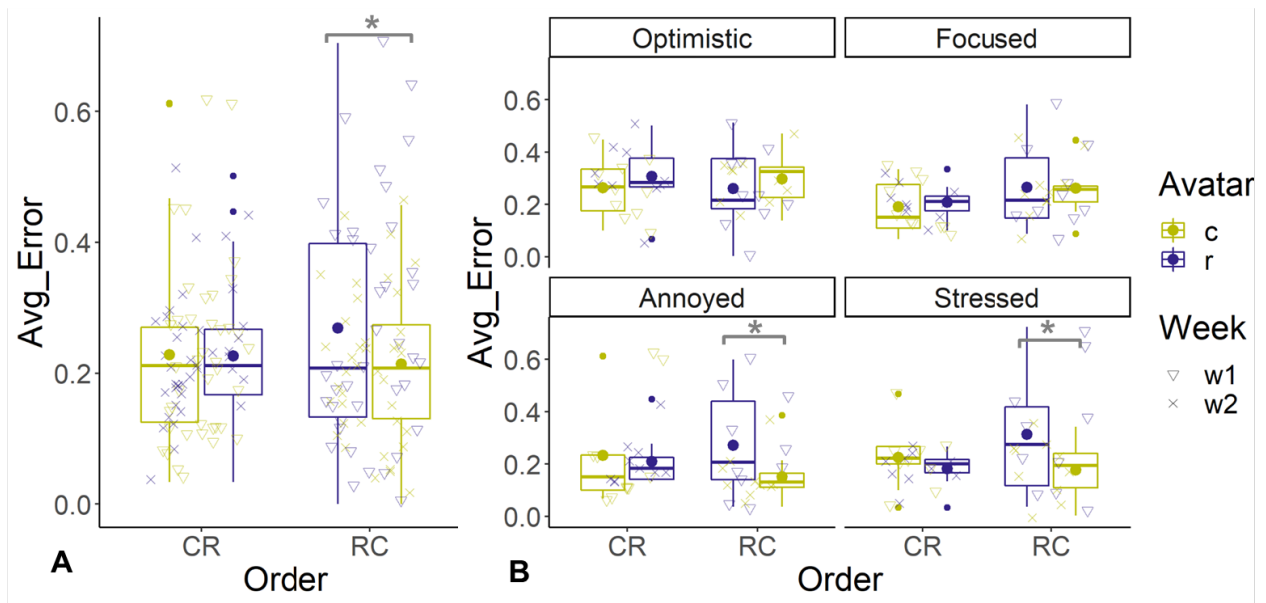


FIGURE 5.5: Averaged error of the perceived emotional state by Order

Boxplots showing the averaged mapping error of the perceived emotional state 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 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 style is separated by order and by emotional state, with each boxplot showing the error of perceiving a specific emotional state (Optimistic, Focused, Annoyed, or Stressed), separated by avatar style.

RQ5b: Accuracy of perceived emotional states over time

We were interested in checking whether people got better at perceiving their colleagues' emotional states over time. To test this, we used the normalised error of the perceived emotional state as presented in Section 5.5.2. 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 over time for each emotional state for all 10 sessions over the two weeks, maintaining chronological order. There was no significant trend for any of the emotional states (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 calculated the regression overall 10 sessions separated by order. There was no significant result for the Cartoon—Realistic order (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 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 ($R^2 < .00, F(1, 86) = .01, p = .92$). Figure 5.6A shows the trends for the Cartoon—Realistic and Realistic—Cartoon orders.

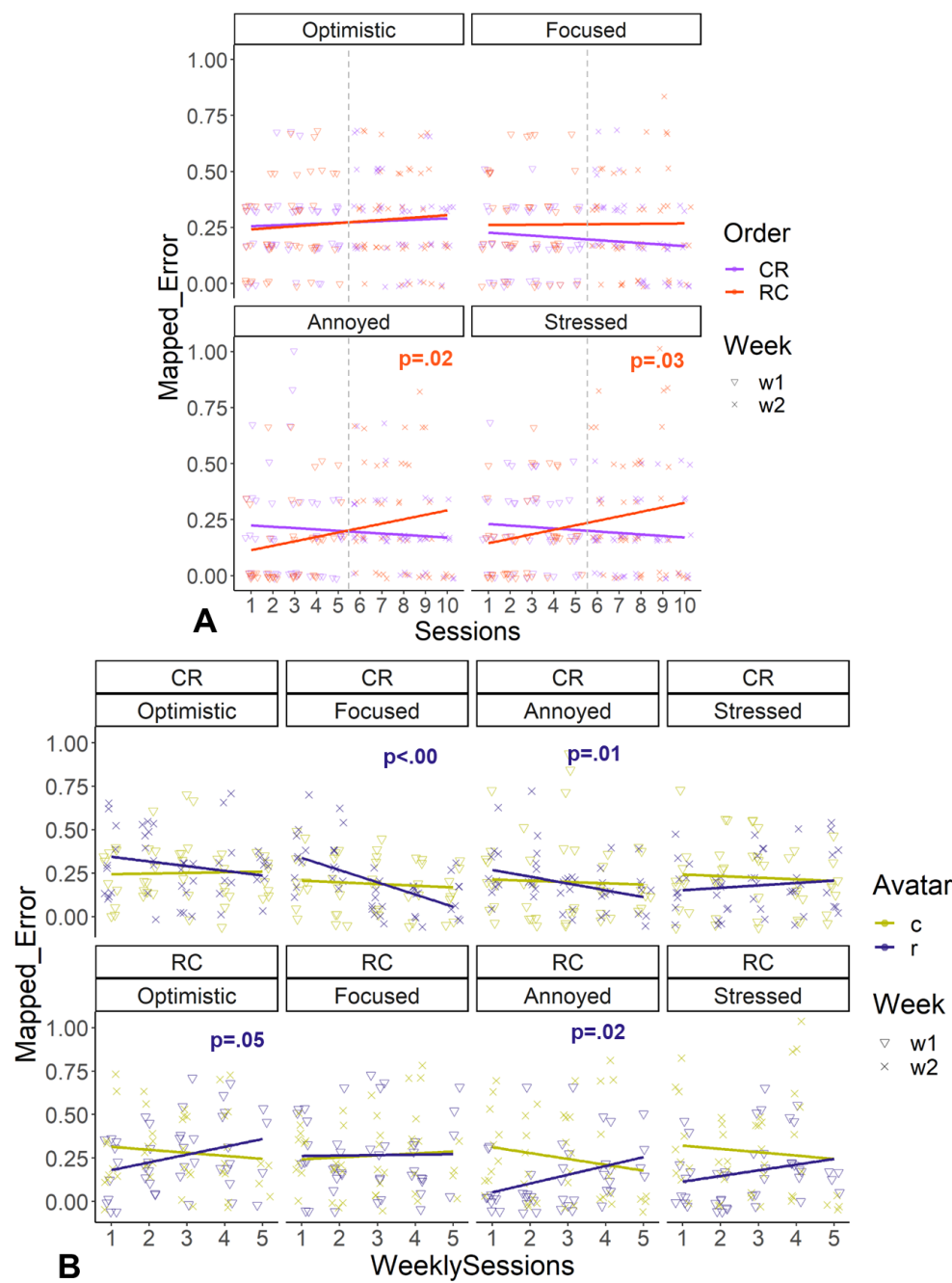


FIGURE 5.6: Error of the perceived emotional state over time

Scatter plots of the average mapping error of perceived emotional state from each participant to each of their colleagues. The mapping error on Y-axis ranges from 0 (no error) to 1 (maximum error). **Panel A:** The X-axis shows the sessions in chronologically order (W1 followed by W2), in Cartoon—Realistic order or Realistic—Cartoon order. The grey dotted lines separate the plots W1 and W2. They are further separated by the emotional states. **B:** The X-axis represents the sessions each group had during one week (1 to 5). The data is separated into CR and RC order and then into the emotional states. Trend lines are fitted for each order and emotional state to show the errors for each avatar style.

Next, we separated the data by avatar style. We calculated the regression for each avatar style considering the order they were used in the weekly meetings, each of which had data from 5 meetings (see Figure 5.6B). 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 ($R^2 = .086, F(1, 42) = 3.96, p=.05$) and Annoyed ($R^2 = .11, F(1, 42) = 5.48, p=.02$). However, for those participants who did not use Realistic avatars until their second week (W2) (after using Cartoon avatars in W1), there was a decrease in the error for Annoyed ($R^2 = .13, F(1, 44) = 6.43, p=.01$) and Focused ($R^2 = .3, F(1, 44) = 19.91, p<.001$). Results with non-significant p-values can be found in the Appendix D.

RQ6: Self-rated emotional states and co-workers' perception errors

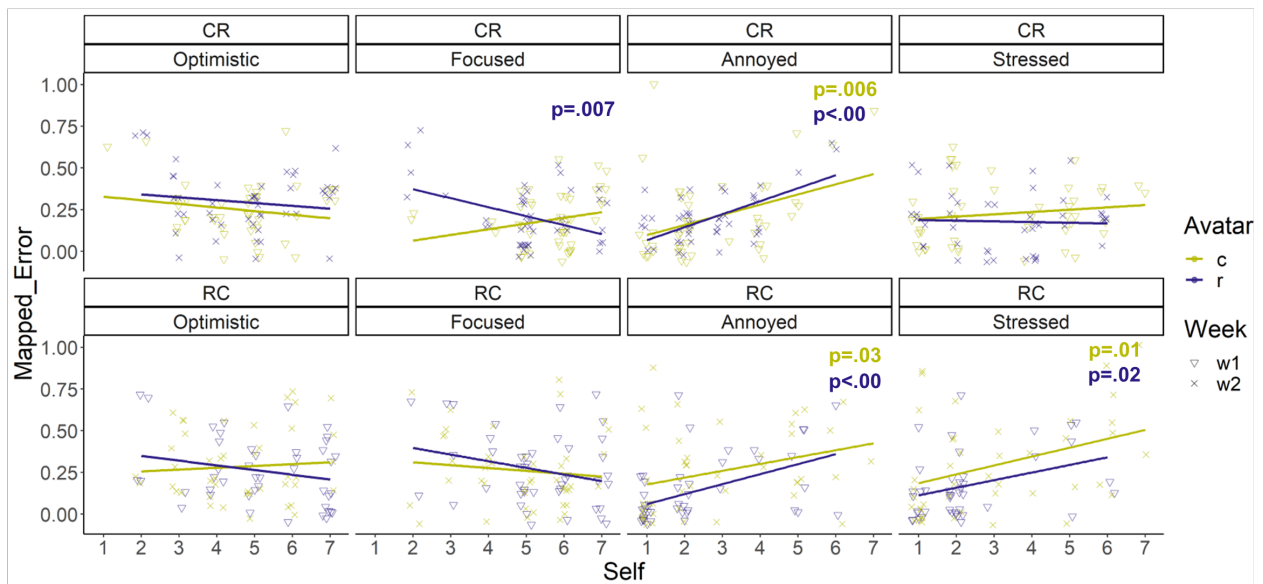


FIGURE 5.7: Error of the perceived emotional state by self-rated emotional states

Scatter plots of mapped error of perceived emotional state 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 emotional state, 1 representing *Strongly Disagree* and 7 representing *Strongly Agree* with having a certain emotional state. The data is separated first by the order (Cartoon—Realistic (top row) and Realistic—Cartoon (bottom row)) and then split by each emotional state (Optimistic, Focused, Annoyed, Stressed), and the data points are fitted for each avatar style (Cartoon and Realistic).

Khojasteh and Won demonstrated a link between high levels of self-reported positive emotions and increased accuracy in perceiving those emotions by others [Khojasteh and Won, 2021]. 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 emotional state and the self-reported ratings of participants' emotional states. We conducted regression analyses for each avatar and for each week (Figure 5.7).

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 ($R^2 = .15, F(1, 44) = 7.99, p = .007$). 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 in perceiving it. When the self-reported ratings were low, their colleagues' perceptions of that emotion had fewer errors. This result was significant for Cartoon avatars (W1: $R^2 = .15, F(1, 44) = 8.11, p = .006$, W2: $R^2 = .09, F(1, 42) = 4.6, p = .03$) and Realistic avatars (W1: $R^2 = .24, F(1, 42) = 13.53, p < .001$, W2: $R^2 = .39, F(1, 44) = 28.8, p < .001$).

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

RQ7: Most useful emotional cues

Participants were asked to rank the usefulness of various cues in perceiving the emotional states 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 Figure 5.8).

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, 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).

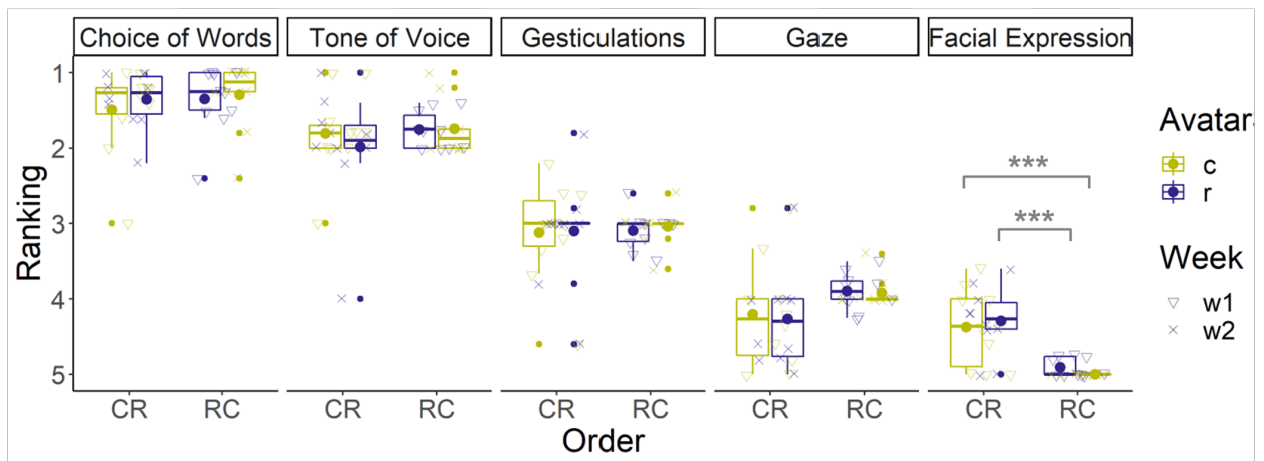


FIGURE 5.8: Ranking of the most useful emotional cues

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

5.6 Discussion and Limitations

The appearance of avatars in social interactions appears 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 style appearance of the avatars; and finally, the greater occurrence of errors in perceiving colleagues’ negative emotions. We then comment on the implications that these findings might have 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 ($p = .17$).

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 the Realistic avatar 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. Further research is needed to separate the emotional effects of avatar appearance from the emotional 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' emotional states (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 a 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' emotional states (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 those 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 emotional states.

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 5.3).

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 5.7 (agree) to 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 emotional states (RQ7, $p < .001$).

Overall, participants using Cartoon avatars made fewer errors in perceiving the emotional states 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 emotional states 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 emotional states (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 emotional states (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 emotional states, potentially leading to greater accuracy.

Use of avatars may lead to more errors perceiving negative emotions

Participants made more errors when trying to perceive their colleagues' negative emotional states. This finding is consistent with the result of Khojasteh and Won [Khojasteh and Won, 2021]. However, this trend was not observed for all emotional states. 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. Finally, there was no trend for errors in perceiving the Optimistic emotional state.

This result might explain some of the outcomes of errors on different avatar styles, for instance, the higher error when perceiving colleagues' emotions when they were using 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 5.5.2), 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 emotional states when using Realistic and Cartoon avatars overall which would explain the higher error rate for perceiving negative emotions in Realistic avatars.

Most of the errors in perceiving emotional states occurred in Realistic W1, so we also compared the self-reported emotional state for each avatar and week. As shown in Section 5.5.2, there was no significant difference between Realistic W1 and Cartoon W1, regardless of the emotional state or when each emotional state was considered separately. There was a difference in the self-reported

emotional states between avatar styles in the Realistic—Cartoon order. Participants in Realistic avatars in their first week self-reported their emotional states 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 emotional states.

Implications

The key issue from the findings above is that, at this stage of reasonable novelty in mixed reality meetings in the workplace, first impressions set expectations but these expectations change over several sessions—for good and for ill. Both realistic and cartoon avatars are useful for establishing co-presence, but cartoon avatars seem to lead to people treating the avatar’s embodiment of *presence* as more important than its ability to visually convey much *nuance of communication or emotion*. The communicative value of realism is currently more fragile than that of cartoon avatars, and it leads people to errors around belief in gaze and emotional perception.

At this point in history, 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, realistic avatars will achieve acceptability within a few years at most, given current advancements in methods for creating and animating realism [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 styles by context (e.g. cartoon for casual and realistic for business, as Mark Zuckerberg believes [Lex Fridman, 2022]). The results of this study suggest that having clear evaluations of how well avatars enable the 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 emotional states, and these errors were different for cartoon versus realistic avatars. These results show that it will be important to disentangle issues of the accuracy of likeness from issues of emotional trustworthiness of likeness.

This brings us to a related point about what it is that 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. 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 videos. Their value will come as a holistic system, and it is in that holism that useful trade-offs in visual realism will be found.

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 results about 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]. Our point is that 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.

Limitations

Although we balanced the participant’s gender, group size and avatar order, we did not take into account the participant’s prior experience with MR devices. Six out of 14 of them never used an 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.

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 nonverbal behaviour gaze or facial expressions. In particular, the results on RQ7 (*What are the most valuable cues available for identifying emotional states and are these different depending on the avatar style*) might have been different if more detailed gaze and facial cues had been implemented.

5.7 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 in MR while embodying two different avatar 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 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 don't live up to expectations for other cues. A crucial finding was that avatars may be less effective in 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.

6

Conclusion & Future Work

In this chapter we revisit the main research questions detailed in Chapter 1, we summarise the contributions from Chapters 3, 4 and 5. Then we discuss the limitations of these works and their impact, and finally, we cover some areas of promising further research in this field of nonverbal behaviours for VCs during social interactions in IVEs.

6.1 Summarised contributions

In this thesis, we introduced work on the nonverbal cues during social interactions in immersive virtual environments covering three areas of interest: People, Agents, and Avatars. Our aim was to move the field forward toward building autonomous agents. Given the circumstances of a doctorate thesis, building a fully autonomous agent was not feasible. Instead, we investigated three main areas regarding virtual characters and their non-verbal cues during social interactions in IVEs. At the same time, we motivated our work based on the industry demands. We worked with different stakeholders to learn their needs and how we can perform research focused on their requirements.

First, in our study on **People** from Chapter 3, we collaborated with the Institute of Cognitive Neuroscience at the University College London (UCL) to understand the dynamics of conversational cues during social interactions in dyads. In this work, we were interested in understanding the dynamics of social interactions between people. Our RQ1 was **What are the dynamics of low-level non-verbal behaviours, in particular gaze and turn-taking, in face-to-face social interactions between two people?**. We developed this further and broke it down into a few hypotheses. First, we wanted to check if a well-established outcome from the literature stands in our dataset: *listeners perform more direct gaze than speakers*. This was our first hypothesis. Further, we were interested in the conversational dynamics and the approach/avoidance conflict introduced by Argyle and Dean [Argyle and Dean, 1965]. Hence our second hypothesis that we

split two parts was: *when someone is being looked at (receiving direct gaze), they would switch back and forth between performing direct gaze and avert gaze with a higher frequency compared to when they are not being looked at*; and the second part is related to when people are not being looked at: *when someone is not being looked at (receiving avert gaze), they would look more at the other person's face (performing direct gaze) than somewhere else- performing avert gaze*. To test these hypotheses, we made use of a multimodal dataset from dyads performing unstructured tasks. This dataset was recorded by the Institute of Cognitive Neuroscience at UCL. We developed the hypothesis, analysed the data, and built a model to automatically annotate speech and gaze targets in dyads. Given that we used an automatic way of annotating the dataset, we used the first hypothesis as a validation of our method of automatically annotating the dataset. In terms of conversational dynamics, we found out that people have a higher frequency of gaze change (from advert to direct and vice versa) when they are being looked at compared to when they are not. When a participant is not being looked at, they also tend to look more at their conversation partner compared to when they are being looked at.

After gaining more insights into conversational dynamics and how complex the social interactions between people are, we decided to focus on higher-level nonverbal behaviours. These tend to be harder to describe using rules and even to explain in detail. They are playing an important role when it comes to building the behaviour of a virtual agent, especially in IVEs, as the plausibility illusion affects greatly the user's experience in this medium (see Section 2.2). Hence, for our second study on **Agents**, we collaborated with two game studios (Dream Reality Interaction and Maze Theory) and we focused on higher-level nonverbal behaviours. For this, we developed the RQ2: **How can an agent be trained to recognise implicit social attitudes during social interactions in virtual reality?** Our industry collaborators were interested in how to recognise implicit social attitudes as this ability would allow for more engaging and inclusive game mechanics in VR games, especially in narrative VR games. To do this, we took into account the gamers' behaviour in a game, and the game creators' creative process and we made use of the technology available to increase/fulfil the market reach need for the game production. We build an immersive ML pipeline to recognise implicit social attitudes in interactions. This pipeline was designed to be used by game designers/creators to annotate abstract behaviours in immersive environments (VR) and eventually to design new interactions in this medium by having an agent recognise implicit non-verbal behaviours. We focused on recognising social engagement and worked on agents (or non-player characters) for the Peaky Blinders narrative VR game developed by Maze Theory. The pipeline built allowed game creators to record and annotate social interactions between an agent and a player, in the same medium that the game will be in (VR). This allowed creators to design more engaging experiences in IVEs. The technical contribution of this work lays in designing the architecture of the ML model. We train a reinforcement machine learning algorithm with imitation learning rewards using raw data (e.g. head position) and socially meaningful derived data (e.g. proxemics); we compared different ML configurations including pre-training and a temporal memory (long-short term memory model - LSTM). The pre-training and LSTM configuration using derived data performed the best (84% F1-score, 83% accuracy) whereas the model using raw data did not generalise.

Hence, addressing our research question, we trained an ML model to recognise implicit social attitudes during social interaction in VR. Although we run our analysis on a specific social attitude (social engagement), we argue that the pipeline can be generalised (see Section 4.6) and it was implemented by the Dream Reality Interactions (one of the industry partners) in a VR karaoke prototype. Our pipeline relies on immersive data annotation without the need to formally define the social attitude. Further, the ML model configuration includes the synchronised data from all participants in an interaction (in this case, the user and the virtual character), a pre-trained reinforcement ML model with imitation learning rewards, a temporal component (to account for past events in the interaction) and socially meaningful derived data to train the model on.

For our final work, we collaborated with a different industry covering another area of virtual characters' social interaction in IVEs. Collaborative remote work has gained more attention, especially during the pandemic years and the IVEs facilitate and enhance remote interactions. Usually, in IVEs users are represented by their avatars and it is not well understood how their appearance influences their interactions. Given this, we worked closely with the Future of Work department from Microsoft Research Cambridge to study this. The main research question on this work was **RQ3: what is the influence and response to the others' personalised avatars appearance during repeated social interactions between co-workers in mixed reality?** In detail, we were interested in understanding the impact the avatars' appearances have on the functional communicative value, task satisfaction, presence, emotional state perception, and the usefulness of cues during real work meetings in MR. We considered two avatar styles: realistic and cartoon. We asked groups of volunteers to run their usual work meetings in MR for 2 – 3 weeks while embodying full-body personalised realistic or cartoon avatars. Based on daily questionnaire responses, we identified that participants who used realistic avatars first may have had higher expectations and more errors in perceiving their colleagues' emotional states. Those using cartoon avatars first 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' emotions, regardless of avatar style. Results from this work also suggested that participants had more errors when perceiving negative emotional states in their colleagues, with this trend appearing for most emotional states (in line with previous research) but it depended on the avatar style order. This work had shown that the users' appearance in MR meetings matters, in particular, if we look at the order the participants are using the avatar styles. Further, the novelty of an MR device wanes off and the overall impressions change after several sessions.

6.2 Collaborative work

All three projects in this thesis are driven from *collaborative work* with different bodies and different stakeholders for each. For the first project on *people*, the contributions of the Institute of Cognitive Neuroscience at UCL were on the data collection and study design. The main contribution presented in this thesis is the data analysis and the automatic cues annotation (gaze and speech). This project helped us better understand the complexity of human-human interactions and the interdisciplinary

work enriched our view on building virtual characters for IVEs that are able to perform verbally and non-verbally as it is expected by the user. When we started the work on agents, we collaborated closely with two game studios (Dream Reality Interaction and Maze Theory) and the project was funded by the InnovateUK grant. The direction of the project was motivated by Maze Theory's new game, *Peaky Blinders* and the way the characters in the game (NPCs or non-player characters) should be created in a VR narrative game (see more details in Section 2.4.1). From the very beginning, we met regularly to design the project in a way that would be the most beneficial for both game companies. Finally, for the third project presented here, we worked together with the Microsoft Research lab in Cambridge under the Future of Work team and with the Mixed Reality team. As remote meetings have become more and more part of our routines, we researched the implications of different avatar styles in remote meetings in VR. The appearance of VCs influences the overall social interactions and the users' expectations. The study design and data collection took place at the Microsoft Research lab during a summer internship, the participants being from different teams within the Microsoft corporation. This allowed using real data in our study to strengthen the ecological validity. The data analysis was performed in close collaboration with the Microsoft Research team ensuring that the outcomes would have a direct impact on the remote meetings strategies. The study was also included in the Microsoft New Future of Work Report 2022 [Teevan et al., 2022].

All of these projects were designed such that the outcomes would be directly applied to the stakeholder's game and application. This ensured the decisions in the work were very well-motivated and relevant to the overall outcome in the bigger picture in the industry.

6.3 Limitations and future work

Although the thesis offers valuable insights on VC for social interactions in IVEs, there are still a number of aspects outside the scope of this thesis. Therefore, a few limitations should be noted in this work. These limitations could act as the motivation for further research that could add to the area of VCs in IVEs alongside the work presented here.

To begin, our first study from Chapter 3 has a number of limitations worth mentioning. The multimodal data collected by the Institute of Cognitive Neuroscience at UCL included dyads of people who didn't know each other. The social dynamics cues between people can differ based on the familiarity of the conversation partners. Given this, our result might not generalise to different groups of people knowing each other. Further work is needed here to test whether these results stand. The data used in this study is multimodal including upper-body motion capture and rich eye data. Our work focused on speech turns and direct/averted gaze direction (looking at the other person or not looking at the other person). More complex data could have been used, however, it was outside of the scope of this work. We believe that future work examining other social dynamics in free-flow conversation could bring important contributions to the field and could add to the insights presented in this thesis. Finally, although we proposed and used a novel automatic data annotation method for this data, this might be seen as a limitation. Our first hypothesis validated

the automatically annotated data, however, a replication of this study using manual annotations could strengthen our results.

The results replication of the second study from Chapter 4 might pose a difficulty without the same annotator. This might seem like a limitation of our work, however, the aim was not to develop a generalisable model for engagement detection (or any other social attitudes). We argue that these abstract social attitudes are hard to define and there are many definitions for engagement [Glas and Pelachaud, 2015]. Thus, our method allows the annotator (in our case the creative game designer) to decide which behaviour can be categorised as social engagement by watching a playback of the interaction in the same medium as it was recorded. This data is then used to train the ML model and when real-life players exhibited the annotated behaviours during gameplay, certain events (agent behaviours, or change of game environment) could be then triggered. Our study presented a way to allow a VC to sense and recognise the events that happen in social interactions in order to respond accordingly. However, the responding part was out of the scope of our work. A potentially fruitful future work would be to implement a response mechanism based on the sensing and recognising described in this thesis. This could create the basis of an autonomous virtual agent that is able to recognise abstract implicit cues in social interactions in IVEs, allowing for novel input interactions, game mechanics or behaviours based on the user-VC rapport.

Our final study on Avatars from Chapter 5 brought valuable insights into how the avatars' appearance impacts social interactions. Aspects such as the longitudinal data collection, the use of personalised full-body avatars or the nature of the task (real-world meetings between co-workers who know each other) strengthen the ecological validity of the study. However, there are some limitations that should be taken into account in future works. First, our results are based on the participants' self-report by filling in questionnaires after each work session in MR. Other data could have also been considered, such as gaze targets, speech or body movements. Furthermore, the avatars used did not have implemented facial expressions nor user-synchronised blinks and lips movement (the blinks were generated and for the lips, we used a general lip-flapping algorithm based on the voice amplitude). Due to time limitations, we could not implement these. Further research is needed to confirm and move our work forward in this fast-moving field.

The participants from all three studies in this thesis come mostly from the western part of the world. In the first two studies (*People* and *Agents*) all participants were volunteers from London, UK; in the last *Avatars* study, the participants were Microsoft Corporation employees from the UK, US, and Africa. Although there are more participants diversity in the final study, it is likely that the ones who volunteer might have had an interest in VR/MR. This is a limitation of all three studies and further work based on people from different cultures could enrich the data collection and would make a valuable contribution to the field.

6.4 Outlook

In this section, I reflect on the work presented in this thesis and on the course of these research areas in the past half a decade.

This thesis contributes to the area of virtual characters in IVEs in three different fields. The initial PhD proposal was more technical covering mostly algorithms' improvements, whilst still focusing on interdisciplinary topics. I thought that I would specialise in certain types of algorithms to create nonverbal behaviours for (autonomous) virtual agents. Right from the first years, I spent most of the time getting up to date with the literature and foundations of social interactions and non-verbal cues. This was very beneficial as it helped me see my work as part of a larger picture. A big part of my PhD work was not coding-heavy as I thought it would be. But rather I focused on the literature from many research fields, which let me understand and reflect on how my technical skills could fit in and help advance the area.

I did not expect my work to be as interdisciplinary, or with such direct applicability in the industry. Working with different stakeholders made me realise that the research needs to be adapted to the final product/application. Working with another academic body led to important contributions to knowledge in form of validating current results and discovering other cues' dynamics in social interactions. These are not necessarily directly applicable to a product but they represent building blocks towards developing virtual agents.

The collaboration with the two industries shaped differently. Firstly, working with two game companies meant opening to adapting usual research practices to fit the project's aim. For instance, there are many virtual characters that are mostly listeners in interactions (e.g., in public speaking, interviewing or direction/information-seeking applications). However, in games it is unusual to have players actively speak to virtual characters (NPCs), hence the virtual characters are the active speakers and the players take mainly the listening role. Thus, the project had different constraints from the technologies used, to the study design. Though, this allowed accommodating a direct application of our results and practices into the game studios' workflows.

The experience was different when working with another industry more focused on research and within a big corporation (Microsoft Research). The project's outcomes were planned to be applied to existing products, tackling fast-moving fields. The remote/hybrid working sector had a large shift and interest in the last years. The primary goals were researching cutting-edge technologies and planning to contribute and advance an existing product.

Looking back, the research projects from this thesis and throughout the PhD have contributed to advances in an interdisciplinary field of autonomous agents in IVEs, but at the same time, have made an impact in two industries covering entertainment and remote work. Looking forward, I am optimistic I will come across more projects with applied research contributions in the areas of nonverbal behaviours for socially interactive autonomous agents.

6.5 Conclusion

This manuscript explored the field of Virtual Characters in Immersive Virtual Environments focusing on nonverbal behaviours in social interactions. We were motivated to build an autonomous virtual character that would be able to socially interact with users in immersive virtual environments without weakening the user’s plausibility illusion.

Throughout the PhD, we have collaborated with different institutions and companies to gather insights from different related fields (as this work is interdisciplinary), and to ensure that the research outcomes are directly applicable in the *real world*. This allowed designing studies with strong ecological validity that can be replicated and expanded through future research.

Thus, this thesis is centred on three main studies. First, we collaborated with the Institute of Cognitive Neuroscience from University College London and using multimodal data between two **people**, we gathered insights into the dynamics of low-level non-verbal behaviours, particularly on gaze targets and turn-taking.

We then considered higher-level nonverbal behaviours. We collaborated with two game studios, Dream Reality Interaction and Maze Theory to research how can an agent be trained to recognise implicit social attitudes. We build an ML model that could be embedded in **agents** to sense and recognise implicit social attitudes when interacting with players with an accuracy of 83%. This was motivated by the need for virtual agents (non-player characters) in Maze Theory’s narrative VR game *Peaky Blinders*.

Lastly, we collaborated with the Future of Work lab and the Mixed Reality lab from Microsoft Research Cambridge to research on **avatars** in remote meetings. We were interested in what is the influence and response that people would have to the others’ personalised avatars’ appearance during repeated meetings between co-workers in MR. We created full-body personalised avatars for all participants in two styles: cartoon and realistic. We compared their experience of using these avatars for 2 – 3 weeks during their usual work meeting in MR. The results imply the use of realistic avatars first, increases the participants’ expectations and they make more errors in perceiving their colleagues’ emotional states. When it comes to cartoon avatars, they may also become more accustomed as time passes.

This thesis offers a number of findings relevant to creating autonomous agents for immersive virtual environments, with a focus on nonverbal behaviour during social interactions with a user. Moreover, the contributions are tangible to relevant fields through our collaboration with the games and work industries.

Bibliography

- [Ahuja et al., 2019] Ahuja, C., Ma, S., Morency, L.-P., and Sheikh, Y. (2019). To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84.
- [Al Moubayed and Skantze, 2011] Al Moubayed, S. and Skantze, G. (2011). Turn-taking control using gaze in multiparty human-computer dialogue: Effects of 2d and 3d displays. In *International Conference on Audio-Visual Speech Processing 2011, Aug 31-Sep 3 2011, Volterra, Italy*, pages 99–102. KTH Royal Institute of Technology.
- [Andrist et al., 2017] Andrist, S., Gleicher, M., and Mutlu, B. (2017). Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*.
- [Andrist et al., 2012] Andrist, S., Pejsa, T., Mutlu, B., and Gleicher, M. (2012). Designing effective gaze mechanisms for virtual agents. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 705.
- [Argyle and Dean, 1965] Argyle, M. and Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, pages 289–304.
- [Argyle and Ingham, 1972a] Argyle, M. and Ingham, R. (1972a). Gaze, Mutual Gaze, and Proximity. *Semiotica*.
- [Argyle and Ingham, 1972b] Argyle, M. and Ingham, R. (1972b). Gaze, mutual gaze, and proximity. *Semiotica*, 6(1):32–49.
- [Aseeri and Interrante, 2021] Aseeri, S. and Interrante, V. (2021). The influence of avatar representation on interpersonal communication in virtual social environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2608–2617.
- [Bailenson and Beall, 2006] Bailenson, J. N. and Beall, A. C. (2006). Transformed Social Interaction: Exploring the Digital Plasticity of Avatars. In Schroeder, R. and Axelsson, A.-S., editors, *Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments*, pages 1–16. Springer Netherlands, Dordrecht.
- [Bailenson and Blascovich, 2004] Bailenson, J. N. and Blascovich, J. (2004). Avatars. encyclopedia of human-computer interaction. *Berkshire Publishing Group*, 64:68.

- [Bailenson et al., 2001] Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments*, 10(6):583–598.
- [Bailenson et al., 2003] Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and social psychology bulletin*, 29(7):819–833.
- [Bailenson and Yee, 2006] Bailenson, J. N. and 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 and Virtual Environments*, 15(6):699–716.
- [Baniyadi et al., 2020] Baniyadi, T., Ayyoubzadeh, S. M., and Mohammadzadeh, N. (2020). Challenges and practical considerations in applying virtual reality in medical education and treatment. *Oman medical journal*, 35(3):e125.
- [Baron-Cohen et al., 1997] Baron-Cohen, S., Wheelwright, S., Jolliffe, and Therese (1997). Is there a "language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. *Visual cognition*, 4(3):311–331.
- [Bayliss et al., 2006] Bayliss, A. P., Paul, M. A., Cannon, P. R., and Tipper, S. P. (2006). Gaze cuing and affective judgments of objects: I like what you look at.
- [Bee et al., 2009] Bee, N., Franke, S., and André, E. (2009). Relations between facial display, eye gaze and head tilt: Dominance perception variations of virtual agents. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE.
- [Bellanca et al., 2019] Bellanca, J. L., Orr, T. J., Helfrich, W. J., Macdonald, B., Navoyski, J., and Demich, B. (2019). Developing a virtual reality environment for mining research. *Mining, metallurgy & exploration*, 36(4):597–606.
- [Benford et al., 1995] Benford, S., Bowers, J., Fahlén, L. E., Greenhalgh, C., and Snowdon, D. (1995). User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 242–249.
- [Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- [Blomqvist and Cook, 2018] Blomqvist, K. and Cook, K. S. (2018). Swift trust: State-of-the-art and future research directions. *The Routledge companion to trust*, pages 29–49.
- [Bohus and Horvitz, 2010] Bohus, D. and Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*.

- [Bohus and Horvitz, 2014] Bohus, D. and Horvitz, E. (2014). Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 2–9. Association for Computing Machinery.
- [Bradski, 2000] Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123.
- [Brugel et al., 2015] Brugel, S., Postma-Nilsenová, M., and Tates, K. (2015). The link between perception of clinical empathy and nonverbal behavior: The effect of a doctor's gaze and body orientation. *Patient Education and Counseling*, 98(10):1260 – 1265. *Communication in Healthcare: Best papers from the International Conference on Communication in Healthcare*, Amsterdam, The Netherlands, 28 September - 1 October 2014.
- [Burgoon et al., 2006] Burgoon, J., Dillman, L., and Stem, L. (2006). Adaptation in dyadic interaction: Defining and operationalizing patterns of reciprocity and compensation. *Communication Theory*, 3:295 – 316.
- [Burgoon et al., 2016] Burgoon, J. K., Guerrero, L. K., and Floyd, K. (2016). *Nonverbal communication*. Routledge.
- [Burke et al., 2016] Burke, S. L., Bresnahan, T., Li, T., Epnere, K., Rizzo, A., Partin, M., Ahlness, R. M., and Trimmer, M. (2016). Using Virtual Interactive Training Agents (ViTA) with Adults with Autism and Other Developmental Disabilities.
- [Cafaro et al., 2016] Cafaro, A., Ravenet, B., Ochs, M., Vilhjálmsón, H. H., and Pelachaud, C. (2016). The effects of interpersonal attitude of a group of agents on user's presence and proxemics behavior. *ACM Trans. Interact. Intell. Syst.*, 6(2).
- [Cañigüeral et al., 2021] Cañigüeral, R., Ward, J. A., and Hamilton, A. F. d. C. (2021). Effects of being watched on eye gaze and facial displays of typical and autistic individuals during conversation. *Autism*, 25(1):210–226.
- [Chinchor, 1992] Chinchor, N. (1992). Muc-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, page 22–29, USA. Association for Computational Linguistics.
- [Chiu et al., 2015] Chiu, C.-C., Morency, L.-P., and Marsella, S. (2015). Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer.
- [Christensen et al., 2018] Christensen, J. V., Mathiesen, M., Poulsen, J. H., Ustrup, E. E., and Kraus, M. (2018). Player experience in a vr and non-vr multiplayer game. In *Proceedings of the Virtual Reality International Conference-Laval Virtual*, pages 1–4. Association for Computing Machinery.
- [Clark, 1996] Clark, H. H. (1996). *Using Language*.
- [Clark and Brennan, 1991] Clark, H. H. and Brennan, S. E. (1991). *Grounding in communication. Perspectives on socially shared cognition*.

- [Clark and Carlson, 1982] Clark, H. H. and Carlson, T. B. (1982). Hearers and speech acts. *Language*.
- [Clark and Marshall, 1981] Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. *Elements of discourse understanding*.
- [Clark and Schaefer, 1987] Clark, H. H. and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*.
- [Clark and Schaefer, 1989] Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*.
- [Clark and Wilkes-Gibbs, 1986] Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- [Collingwoode-Williams et al., 2021] Collingwoode-Williams, T., O’Shea, Z., Gillies, M., and Pan, X. (2021). The impact of self-representation and consistency in collaborative virtual environments. *Frontiers in Virtual Reality*, 2:648–601.
- [Cummins, 2012] Cummins, F. (2012). Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes*.
- [De Simone et al., 2019] De Simone, F., Li, J., Debarba, H. G., El Ali, A., Gunkel, S. N., and 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)*, pages 890–891. IEEE.
- [Dermouche and Pelachaud, 2019a] Dermouche, S. and Pelachaud, C. (2019a). Engagement modeling in dyadic interaction. In *2019 International Conference on Multimodal Interaction*, pages 440–445. Association for Computing Machinery.
- [Dermouche and Pelachaud, 2019b] Dermouche, S. and Pelachaud, C. (2019b). Generative model of agent’s behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*, pages 375–384. Association for Computing Machinery.
- [DeVault et al., 2014] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-p. (2014). SimSensei Kiosk : A virtual human interviewer for healthcare decision support. *International Conference on Autonomous Agents and Multi-Agent Systems*.
- [Dhamija and Boulton, 2017] Dhamija, S. and Boulton, T. E. (2017). Automated mood-aware engagement prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

- [Dhananjaya and Yegnanarayana, 2008] Dhananjaya, N. G. and Yegnanarayana, B. (2008). Speaker change detection in casual conversations using excitation source features. *Speech communication*, 50(2):153–161.
- [Dias et al., 2014] Dias, J., Mascarenhas, S., and Paiva, A. (2014). Fatima modular: Towards an agent architecture with a generic appraisal framework. *Emotion modeling: Towards pragmatic computational models of affective processes*, pages 44–56.
- [Duncan, 1972] Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*.
- [Duncan and Fiske, 1977] Duncan, S. and Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory*.
- [Efran and Broughton, 1966] Efran, J. S. and Broughton, A. (1966). Effect of expectancies for social approval on visual behavior. *Journal of Personality and Social Psychology*, 4(1):103.
- [Ellyson et al., 1981] Ellyson, S. L., Dovidio, J. F., and Fehr, B. (1981). Visual behavior and dominance in women and men. In *Gender and nonverbal behavior*, pages 63–79. Springer.
- [Emery, 2000] Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604.
- [Exline, 1963] Exline, R. V. (1963). Explorations in the process of person perception: visual interaction in relation to competition, sex, and need for affiliation. *Journal of personality*, 31(1):1–20.
- [Félix-Brasdefer, 2015] Félix-Brasdefer, J. C. (2015). *The language of service encounters*. Cambridge University Press.
- [Feng et al., 2017] Feng, W., Kannan, A., Gkioxari, G., and Zitnick, C. L. (2017). Learn2smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4131–4138.
- [Ferstl and McDonnell, 2018] Ferstl, Y. and McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98. ACM.
- [Fischer and Tenbrink, 2003] Fischer, K. and Tenbrink, T. (2003). Video conferencing in a transregional research cooperation: Turn-taking in a new medium. *Connecting Perspectives, Videokonferenz: Beiträge zu ihrer Erforschung und Anwendung*, Aachen: Shaker Verlag.
- [Forbes-Riley et al., 2012] Forbes-Riley, K., Litman, D., Friedberg, H., and Drummond, J. (2012). Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 91–102, Montréal, Canada. Association for Computational Linguistics.
- [Freeth et al., 2013] Freeth, M., Foulsham, T., and Kingstone, A. (2013). What affects social attention? social presence, eye contact and autistic traits. *PloS one*, 8(1):e53286.

- [Freiwald et al., 2021] Freiwald, J. P., Schenke, J., Lehmann-Willenbrock, N., and Steinicke, F. (2021). Effects of avatar appearance and locomotion on co-presence in virtual reality collaborations. In *Mensch und Computer 2021*, pages 393–401.
- [Galesic, 2006] Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of official statistics*, 22(2):313.
- [Garcia et al., 2021] Garcia, B., Chun, S., Kicklighter, C., Mai, B., Palma, M., and Seo, J. H. (2021). Studying design attributes of virtual characters to support students’ perceived experiences in virtual reality lectures. *International Association for Development of the Information Society*.
- [Garrod and Pickering, 2004] Garrod, S. and Pickering, M. J. (2004). Why is conversation so easy?
- [Gillies et al., 2015] Gillies, M., Kleinsmith, A., and Brenton, H. (2015). Applying the CASSM framework to improving end user debugging of interactive machine learning. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume 2015-Janua.
- [Glas and Pelachaud, 2015] Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 944–949. IEEE.
- [Goodwin, 1981] Goodwin, C. (1981). Conversational Organization: Interaction Between Speakers and Hearers. *Conversational Organization: Interaction Between Speakers and Hearers*.
- [Goodwin, 1986] Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*.
- [Gordon et al., 2019] Gordon, C., Leuski, A., Benn, G., Klassen, E., Fast, E., Liewer, M., Hartholt, A., and Traum, D. R. (2019). Primer: An emotionally aware virtual agent. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI2019), Los Angeles, USA, March 20, 2019*, volume 2327 of *CEUR Workshop Proceedings*.
- [Gordon et al., 2016] Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., and Breazeal, C. (2016). Affective personalization of a social robot tutor for children’s second language skills. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [Greenwood et al., 2017] Greenwood, D., Laycock, S., and Matthews, I. (2017). Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents*, pages 160–169. Springer.
- [Grice, 1975] Grice, H. P. (1975). *Logic and conversation*.
- [Gupta, 2015] Gupta, V. (2015). Speaker change point detection using deep neural nets. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4420–4424. IEEE.

- [Haag and Shimodaira, 2016] Haag, K. and Shimodaira, H. (2016). Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *International Conference on Intelligent Virtual Agents*, pages 198–207. Springer.
- [Hale et al., 2020] Hale, J., Ward, J. A., Buccheri, F., Oliver, D., and Hamilton, A. F. d. C. (2020). Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior*, 44(1):63–83.
- [Hall, 1966] Hall, E. T. (1966). *The hidden dimension*, volume 609. Garden City, NY: Doubleday.
- [Hall et al., 1968] Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannon, 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]. *Current anthropology*, 9(2/3):83–108.
- [Han et al., 2022] Han, E., Miller, M. R., Ram, N., Nowak, K. L., and 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 and Biocca, 2004] Harms, C. and Biocca, F. (2004). Internal consistency and reliability of the networked minds measure of social presence.
- [Hartholt et al., 2022] Hartholt, A., Fast, E., Li, Z., Kim, K., Leeds, A., and Mozgai, S. (2022). Re-architecting the virtual human toolkit: towards an interoperable platform for embodied conversational agent research and development. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8. Association for Computing Machinery.
- [Hartholt et al., 2020] Hartholt, A., Fast, E., Reilly, A., Whitcup, W., Liewer, M., and Mozgai, S. (2020). Multi-platform expansion of the virtual human toolkit: ubiquitous conversational agents. *International Journal of Semantic Computing*, 14(03):315–332.
- [Hartholt et al., 2019] Hartholt, A., Mozgai, S., Fast, E., Liewer, M., Reilly, A., Whitcup, W., and Rizzo, A. S. (2019). Virtual humans in augmented reality: A first step towards real-world embedded virtual roleplayers. In *Proceedings of the 7th international conference on human-agent interaction*, pages 205–207. Association for Computing Machinery.
- [Hartholt et al., 2013] Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L. P., and Gratch, J. (2013). All together now: Introducing the virtual human toolkit. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- [Haykin and Chen, 2005] Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural computation*, 17(9):1875–1902.
- [Heidicker et al., 2017] Heidicker, P., Langbehn, E., and Steinicke, F. (2017). Influence of avatar appearance on presence in social vr. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 233–234. IEEE.
- [Herrera et al., 2020] Herrera, F., Oh, S. Y., and Bailenson, J. N. (2020). Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence*, 27(2):163–182.

- [Hessels et al., 2019] Hessels, R. S., Holleman, G. A., Kingstone, A., Hooge, I. T., and Kemner, C. (2019). Gaze allocation in face-to-face communication is affected primarily by task structure and social context, not stimulus-driven factors. *Cognition*, 184:28–43.
- [Ho and Ermon, 2016] Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573. Curran Associates Inc.
- [Ho et al., 2015] Ho, S., Foulsham, T., and Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Hoegen et al., 2019] Hoegen, R., Aneja, D., McDuff, D., and Czerwinski, M. (2019). An end-to-end conversational style matching agent.
- [Jo et al., 2017] Jo, D., Kim, K.-H., and 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*, pages 27–36.
- [Kendon, 1967] Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*.
- [Kendon and Cook, 1969] Kendon, A. and Cook, M. (1969). The consistency of gaze patterns in social interaction. *British journal of psychology (London, England : 1953)*.
- [Kendrick and Holler, 2017] Kendrick, K. H. and Holler, J. (2017). Gaze Direction Signals Response Preference in Conversation. *Research on Language and Social Interaction*.
- [Kenny et al., 2007] Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D., and Rey, M. D. (2007). Building Interactive Virtual Humans for Training Environments. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2007*.
- [Kenny et al., 2009] Kenny, P. G., Parsons, T. D., and Rizzo, A. A. (2009). Human computer interaction in virtual standardized patient systems. In *International Conference on Human-Computer Interaction*, pages 514–523.
- [Khakhulin et al., 2022] Khakhulin, T., Sklyarova, V., Lempitsky, V., and Zakharov, E. (2022). Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer.
- [Khaki et al., 2016] Khaki, H., Bozkurt, E., and Erzin, E. (2016). Agreement and disagreement classification of dyadic interactions using vocal and gestural cues. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2762–2766. IEEE.
- [Khojasteh and Won, 2021] Khojasteh, N. and Won, A. S. (2021). Working together on diverse tasks: A longitudinal study on individual workload, presence and emotional recognition in collaborative virtual environments. *Frontiers in Virtual Reality*, 2:53.

- [Koch et al., 2018] Koch, M., von Luck, K., Schwarzer, J., and 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).
- [Kotti et al., 2008] Kotti, M., Moschou, V., and Kotropoulos, C. (2008). Speaker segmentation and clustering. *Signal processing*, 88(5):1091–1124.
- [Lane et al., 2011] Lane, H. C., Noren, D., Auerbach, D., Birch, M., and Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- [Langa et al., 2022] Langa, S. F., Montagud, M., Cernigliaro, G., and Rivera, D. R. (2022). Multiparty holomeetings: Toward a new era of low-cost volumetric holographic meetings in virtual reality. *Ieee Access*, 10:81856–81876.
- [Latoschik et al., 2017] Latoschik, M. E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., and 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*, pages 1–10.
- [Le et al., 2012] Le, B. H., Ma, X., and Deng, Z. (2012). Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*.
- [Lee et al., 2002] Lee, S. P., Badler, J. B., and Badler, N. I. (2002). Eyes Alive.
- [Lex Fridman, 2022] Lex Fridman (2022). Mark Zuckerberg: Meta, Facebook, Instagram, and the Metaverse. Feb 26, 2022.
- [Lombard et al., 2009] Lombard, M., Ditton, T. B., and Weinstein, L. (2009). Measuring presence: the temple presence inventory. In *Proceedings of the 12th annual international workshop on presence*.
- [Lugrin et al., 2015] Lugrin, J.-L., Latt, J., and Latoschik, M. E. (2015). Anthropomorphism and illusion of virtual body ownership. In *ICAT-EGVE*, pages 1–8.
- [Ma and Pan, 2022] Ma, F. and Pan, X. (2022). Visual fidelity effects on expressive self-avatar in virtual reality: First impressions matter. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 57–65. IEEE.
- [Ma et al., 2021] Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De la Torre, F., and Sheikh, Y. (2021). Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73.
- [Macdonald and Tatler, 2013] Macdonald, R. G. and Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of vision*, 13(4):6–6.

- [MacDorman and Chattopadhyay, 2016] MacDorman, K. F. and Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, 146:190–205.
- [Marsella et al., 2013] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '13*.
- [Mascarenhas et al., 2018] Mascarenhas, S., Guimarães, M., Prada, R., Dias, J., Santos, P. A., Star, K., Hirsh, B., Spice, E., and Kommeren, R. (2018). A virtual agent toolkit for serious games developers. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–7. IEEE.
- [Mascarenhas et al., 2022] Mascarenhas, S., Guimarães, M., Prada, R., Santos, P. A., Dias, J., and Paiva, A. (2022). Fatima toolkit: Toward an accessible tool for the development of socio-emotional agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(1):1–30.
- [Mason et al., 2005] Mason, M. F., Tatkov, E. P., and Macrae, C. N. (2005). The look of love: Gaze shifts and person perception. *Psychological Science*.
- [Matsuyama et al., 2016] Matsuyama, Y., Bhardwaj, A., Zhao, R., Romero, O. J., Akoju, S. A., and Cassell, J. (2016). Socially-Aware Animated Intelligent Personal Assistant Agent. *SIGDIAL Conference*.
- [Matthews et al., 1990] Matthews, G., Jones, D. M., and Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST Mood Adjective Checklist. *British Journal of Psychology*, 81(1).
- [Max Roser and Ritchie, 2013] Max Roser, C. A. and Ritchie, H. (2013). Human height. *Our World in Data*. <https://ourworldindata.org/human-height>.
- [McLaren et al., 2020] McLaren, L., Koutsombogera, M., and Vogel, C. (2020). A heuristic method for automatic gaze detection in constrained multi-modal dialogue corpora. In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000055–000060. IEEE.
- [Meyerson et al., 1996] Meyerson, D., Weick, K. E., Kramer, R. M., et al. (1996). Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research*, 166:195.
- [Morency et al., 2008] Morency, L.-P., de Kok, I., and Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*, pages 176–190. Springer.
- [Mota and Picard, 2003] Mota, S. R. and Picard, R. W. (2003). Automated posture analysis for detecting learner’s interest level. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, pages 49–49.

- [Moubayed et al., 2012] Moubayed, S. A., Edlund, J., and Beskow, J. (2012). Taming Mona Lisa: Communicating Gaze Faithfully in 2D and 3D Facial Projections. *ACM Transactions on Interactive Intelligent Systems*.
- [Moustafa and Steed, 2018] Moustafa, F. and 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*, pages 1–10.
- [Neff and Pelachaud, 2017] Neff, M. and Pelachaud, C. (2017). Animation of natural virtual characters. *IEEE Computer Graphics and Applications*, 37(4):14–16.
- [Nguyen et al., 2018] Nguyen, D.-C., Bailly, G., and Elisei, F. (2018). Comparing cascaded lstm architectures for generating head motion from speech in task-oriented dialogs. In *International Conference on Human-Computer Interaction*, pages 164–175. Springer.
- [Nielsen, 1964] Nielsen, G. S. (1964). *Studies in self-confrontation*. Munksgaard.
- [Nordin Forsberg and Kirchner, 2021] Nordin Forsberg, B. and Kirchner, K. (2021). The perception of avatars in virtual reality during professional meetings. In *International Conference on Human-Computer Interaction*, pages 290–294. Springer.
- [Oh et al., 2018] Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5.
- [Otto et al., 2006] Otto, O., Roberts, D., and 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*, pages 145–154.
- [Pakanen et al., 2022] Pakanen, M., Alavesä, P., van Berkel, N., Koskela, T., and Ojala, T. (2022). “nice to see you virtually”: Thoughtful design and evaluation of virtual avatar of the other user in ar and vr based teleexistence systems. *Entertainment Computing*, 40:100457.
- [Pan et al., 2018] Pan, X., Collingwoode-Williams, T., Antley, A., Brenton, H., Congdon, B., Drewett, O., Gillies, M. F. P., Swapp, D., Pleasence, P., Fertleman, C., et al. (2018). A study of professional awareness using immersive virtual reality: the responses of general practitioners to child safeguarding concerns. *Frontiers in Robotics and AI*, 5:80.
- [Pan et al., 2012] Pan, X., Gillies, M., Barker, C., Clark, D. M., and Slater, M. (2012). Socially anxious and confident men interact with a forward virtual woman: an experimental study. *PloS one*, 7(4):e32931.
- [Pan and Hamilton, 2018] Pan, X. and Hamilton, A. F. d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109(3):395–417.
- [Pan et al., 2016] Pan, X., Slater, M., Beacco, A., Navarro, X., Rivas, A. I. B., Swapp, D., Hale, J., Forbes, P. A. G., Denvir, C., Hamilton, A. F. d. C., et al. (2016). The responses of medical general practitioners to unreasonable patient demand for antibiotics—a study of medical ethics using immersive virtual reality. *PloS one*, 11(2):e0146837.

- [Pan and Steed, 2017] Pan, Y. and Steed, A. (2017). The impact of self-avatars on trust and collaboration in shared virtual environments. *PloS one*, 12(12):e0189078.
- [Parmar, 2017] Parmar, D. (2017). *Evaluating the effects of immersive embodied interaction on cognition in virtual reality*. PhD thesis, Clemson University.
- [Pauw et al., 2022] Pauw, L. S., Sauter, D. A., van Kleef, G. A., Lucas, G. M., Gratch, J., and Fischer, A. H. (2022). The avatar will see you now: Support from a virtual human provides socio-emotional benefits. *Computers in Human Behavior*, 136:107368.
- [Pejsa et al., 2017] Pejsa, T., Gleicher, M., and Mutlu, B. (2017). Who, me? How virtual agents can shape conversational footing in virtual reality. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- [Pickering and Garrod, 2004] Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*.
- [Recasens et al., 2017] Recasens, A., Vondrick, C., Khosla, A., and Torralba, A. (2017). Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443.
- [Rizzo et al., 2015] Rizzo, A., Cukor, J., Gerardi, M., Alley, S., Reist, C., Roy, M., Rothbaum, B. O., and Difede, J. (2015). Virtual reality exposure for ptsd due to military combat and terrorist attacks. *Journal of Contemporary Psychotherapy*, 45:255–264.
- [Roberts, 2011] Roberts, S. (2011). Dipdap.
- [Sakurai et al., 2021] Sakurai, S., Goto, T., Nojima, T., and Hirota, K. (2021). Effect of the opponent’s appearance on interpersonal cognition that affects user-to-user relationship in virtual whole-body interaction. *Journal of Robotics and Mechatronics*, 33(5):1029–1042.
- [Sandgren et al., 2012] Sandgren, O., Andersson, R., Weijer, J. V. D., Hansson, K., and Sahlén, B. (2012). Timing of gazes in child dialogues: A time-course analysis of requests and back channelling in referential communication. *International Journal of Language and Communication Disorders*.
- [Sanghvi et al., 2011] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction, HRI ’11*, page 305–312, New York, NY, USA. Association for Computing Machinery.
- [Schegloff and Sacks, 1973] Schegloff, E. A. and Sacks, H. (1973). Opening up Closings. *Semiotica*.
- [Schilbach et al., 2013] Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36:393–414.
- [Schmidt, 2000] Schmidt, A. (2000). Implicit human computer interaction through context. *Personal technologies*, 4(2):191–199.

- [Schmuckler, 2001] Schmuckler, M. A. (2001). What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.
- [Schröder, 2010] Schröder, M. (2010). The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction*.
- [Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- [Seele et al., 2017] Seele, S., Misztal, S., Buhler, H., Herpers, R., and Schild, J. (2017). Here’s looking at you anyway! How important is realistic gaze behavior in co-located social virtual reality games? In *CHI PLAY 2017 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*.
- [Shao et al., 2019] Shao, K., Tang, Z., Zhu, Y., Li, N., and Zhao, D. (2019). A survey of deep reinforcement learning in video games.
- [Shin et al., 2019] Shin, M., Kim, S. J., and Biocca, F. (2019). The uncanny valley: No need for any further judgments when an avatar looks eerie. *Computers in Human Behavior*, 94:100–109.
- [Sidenmark and Gellersen, 2019] Sidenmark, L. and Gellersen, H. (2019). Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(1):1–40.
- [Slater, 1999] Slater, M. (1999). Measuring presence: A response to the witmer and singer presence questionnaire. *Presence*, 8(5):560–565.
- [Slater, 2009] Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557.
- [Slater et al., 2000] Slater, M., Sadagic, A., Usoh, M., and Schroeder, R. (2000). Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*.
- [Slater and Sanchez-Vives, 2016] Slater, M. and Sanchez-Vives, M. V. (2016). Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI*.
- [Slater and Steed, 2000] Slater, M. and Steed, A. (2000). A virtual presence counter. *Presence: Teleoperators & Virtual Environments*, 9(5):413–434.
- [Smith and Neff, 2018] Smith, H. J. and Neff, M. (2018). Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.
- [Standaert et al., 2021] Standaert, W., Muylle, S., and Basu, A. (2021). How shall we meet? Understanding the importance of meeting mode capabilities for different meeting objectives. *Information & Management*, 58(1):103393.

- [Stavroulia et al., 2019] Stavroulia, K. E., Christofi, M., Baka, E., Michael-Grigoriou, D., Magnenat-Thalmann, N., and Lanitis, A. (2019). Assessing the emotional impact of virtual reality-based teacher training. *The International Journal of Information and Learning Technology*.
- [Steed and Schroeder, 2015] Steed, A. and Schroeder, R. (2015). Collaboration in immersive and non-immersive virtual environments. In *Immersed in Media*, pages 263–282. Springer.
- [Stratou and Morency, 2017] Stratou, G. and Morency, L.-P. (2017). MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. *IEEE Transactions on Affective Computing*.
- [Sun and Won, 2021] Sun, Y. and Won, A. S. (2021). Despite appearances: Comparing emotion recognition in abstract and humanoid avatars using nonverbal behavior in social virtual reality. *Frontiers in Virtual Reality*, page 109.
- [Swartout et al., 2010] Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J. Y., Gerten, J., Chu, S., and White, K. (2010). Ada and grace: Toward realistic and engaging virtual museum guides. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- [Teevan et al., 2022] Teevan, J., Baym, N., Butler, J., Hecht, B., Jaffe, S., Nowak, K., Sellen, A., Yang, L., Ash, M., Awori, K., Bruch, M., Choudhury, P., Coleman, A., Counts, S., Cupala, S., Czerwinski, M., Doran, E., Fetterolf, E., Gonzalez Franco, M., Gupta, K., Halfaker, A. L., Hadley, C., Houck, B., Inkpen, K., Iqbal, S., Knudsen, E., Levine, S., Lindley, S., Neville, J., O’Neill, J., Pollak, R., Poznanski, V., Rintel, S., Shah, N. P., Suri, S., Troy, A. D., and Wan, M. (2022). Microsoft new future of work report 2022. Technical Report MSR-TR-2022-3, Microsoft.
- [Terada et al., 2021] Terada, K., Okazoe, M., and Gratch, J. (2021). Effect of politeness strategies in dialogue on negotiation outcomes. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 195–202. Association for Computing Machinery.
- [Thiebaux et al., 2008] Thiebaux, M., Marsella, S., Marshall, A. N., and Kallmann, M. (2008). SmartBody: behavior realization for embodied conversational agents. *Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems*.
- [Tompkins, 1963] Tompkins, S. S. (1963). *Affect, imagery, consciousness: II. The Negative Affects*.
- [Traum et al., 2012] Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., and Swartout, W. (2012). Ada and grace: Direct interaction with museum visitors. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- [Tritschler and Gopinath, 1999] Tritschler, A. and Gopinath, R. A. (1999). Improved speaker segmentation and segments clustering using the bayesian information criterion. In *Sixth European Conference on Speech Communication and Technology*.

- [Vinayagamoorthy et al., 2004] Vinayagamoorthy, V., Garau, M., Steed, A., and Slater, M. (2004). An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. *Computer Graphics Forum*.
- [Vinciarelli et al., 2011] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D’Errico, F., and Schroeder, M. (2011). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.
- [Waltemate et al., 2018] Waltemate, T., Gall, D., Roth, D., Botsch, M., and Latoschik, M. E. (2018). The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652.
- [Wang et al., 2017] Wang, R., Gu, M., Li, L., Xu, M., and Zheng, T. F. (2017). Speaker segmentation using deep speaker vectors for fast speaker change scenarios. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5420–5424. IEEE.
- [Woolf et al., 2009] Woolf, B., Bursleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164.
- [Yang et al., 2005] Yang, M., Yang, Y., and Wu, Z. (2005). A pitch-based rapid speech segmentation for speaker indexing. In *Seventh IEEE International Symposium on Multimedia (ISM’05)*, pages 6–pp. IEEE.
- [Yoon et al., 2019] Yoon, B., Kim, H.-i., Lee, G. A., Billinghamurst, M., and 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)*, pages 547–556. IEEE.
- [Yu et al., 2004] Yu, C., Aoki, P. M., and Woodruff, A. (2004). Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027*.
- [Yuan et al., 2019] Yuan, L., Dennis, A., Riemer, K., et al. (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*.
- [Zadeh et al., 2018] Zadeh, A., Lim, Y. C., Baltrušaitis, T., and Morency, L. P. (2018). Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*.
- [Zhang et al., 2022] Zhang, Y., Yang, J., Liu, Z., Wang, R., Chen, G., Tong, X., and Guo, B. (2022). Virtualcube: An immersive 3d video communication system. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2146–2156.
- [Zibrek et al., 2018] Zibrek, K., Kokkinara, E., and McDonnell, R. (2018). The effect of realistic appearance of virtual characters in immersive environments—does the character’s personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690.

- [Zoric et al., 2011] Zoric, G., Forchheimer, R., and Pandzic, I. S. (2011). On creating multimodal virtual humans-real time speech driven facial gesturing. *Multimedia Tools and Applications*.

A

Questionnaires

This appendix chapter contains the questionnaires used for Chapter 4 and Chapter 5.

First, we start with the ones from Chapter 4. It includes the questionnaire after the first VR stage, after the second VR stage and after the third VR stage. After this, there is the Demographic questionnaire.

Next, we present the questionnaires from Chapter 5. It includes the *onboarding questionnaire* followed by the daily questionnaire titled *Daily post-meeting questionnaire*.

All the questionnaires in this appendix are recreations of the questionnaires in the Microsoft Forms online tool. Participants used the more user-friendly web version of Microsoft Forms to answer the questions.

Questionnaire after the first stage

Please complete this questionnaire after the first stage.

1. Li was in the room with you. What do you think about Li?

	Yes	No
I liked Li.		
I would like to spend more time with Li.		
I feel Li is reliable.		
I feel Li is honest.		
I feel Li is sociable.		
I feel Li is friendly.		
I feel Li is sympathetic.		
I feel Li is trustworthy.		

2. Did Li mention any member of her family in her monologue? If so, list the family members you remember.

3. What kind of relationship do you think Li has with her family?

4. How do you think Li was feeling like?

5. What do you think about this session? Please add any comments you might have below.

Please answer the following keeping in mind the most recent experience.

6. Please rate your sense of being in the room with Li, on the following scale from 1 to 7, where 7 represents your normal experience of being in a place.

I had a sense of "being there" in the room with Li:

1 - Not at all	2	3	4	5	6	7 - Very much

7. To what extent were there times during the experience when the virtual room became the "reality" for you, and you almost forgot about the "real world" of the office/lab in which the whole experience was really taking place?

There were times during the experience when the virtual room became more real for me compared to the "real world"...

1 – At no time	2	3	4	5	6	7 – Almost all the time

- 8. When you think back about your experience, do you think of the virtual room/virtual character more as images that you saw, or more as somewhere that you visited? Please answer on the following 1 to 7 scale:**

The virtual room seems to me to be more like...

1 - Images that I saw	2	3	4	5	6	7 - Somewhere that I visited

- 9. During the time of the experience, which was strongest on the whole, your sense of being in the virtual room, or of being in the real world of the lab/office?**

I had a stronger sense of being in...

1 - The real world of the lab/office	2	3	4	5	6	7 - The virtual reality of the room with Li

- 10. Consider your memory of being in the virtual room. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By 'structure of the memory' consider things like the extent to which you have a visual memory of the room, whether that memory is in colour, the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such structural elements.**

I think of the virtual room as a place in a way similar to other places that I've been today....

1 - not at all	2	3	4	5	6	7 - very much so

- 11. During the time of the experience, did you often think to yourself that you were actually just standing in an office wearing a helmet or did the virtual room and Li overwhelm you? During the experience I often thought that I was really standing in the lab/office wearing a helmet....**

1 - most of the time I realised I was in the lab	2	3	4	5	6	7 - never because the virtual room overwhelmed me

Please answer the following keeping in mind the most recent experience.

	1-Not at all	2	3	4	5	6	7-Very much so
How much did you behave within the experience as if the situation was real?							
How much was your emotional response in the experience as if the situation was real?							
How much were you thinking things like 'I know this isn't real' but then surprisingly finding yourself behaving as if it was real?							
To what extent were your physical responses within the experience (e.g. heart rate, blushing sweating, etc) the same as if it had been a real situation?							
How much did you behave as if Li was real?							
How much was your emotional response to Li as if she was real?							
How much were your thoughts in relation to Li as if she was real?							
How much were you thinking things like 'I know this person isn't real' but then surprisingly finding yourself behaving as if Li was real?							
How much did you have physical responses (such as change in heart rate, blushing sweating, etc) towards Li as if she was real?							

Questionnaire after the second stage

1. Please answer the following keeping in mind the most recent experience.

	1-Not at all	2	3	4	5	6	7-Very much so
How much did you behave within the experience as if the situation was real?							
How much was your emotional response in the experience as if the situation was real?							
How much were you thinking things like 'I know this isn't real' but then surprisingly finding yourself behaving as if it was real?							
To what extent were your physical responses within the experience (e.g. heart rate, blushing sweating, etc) the same as if it had been a real situation?							
How much did you behave as if Li was real?							
How much was your emotional response to Li as if she was real?							
How much were your thoughts in relation to Li as if she was real?							
How much were you thinking things like 'I know this person isn't real' but then surprisingly finding yourself behaving as if Li was real?							
How much did you have physical responses (such as change in heart rate, blushing sweating, etc) towards Li as if she was real?							

2. How well would you say you performed at gaining Li's trust?

3. How would you rate the following statement:
Would you say that Li trusts you with delivering an important piece of information to another person?

1 – Definitely no	2	3	4	5	6	7 – Definitely yes

4. What do you think about this, most recent session? Please add any comments you might have below.

Questionnaire after the third stage

1. Please answer the following keeping in mind the most recent experience.

	1-Not at all	2	3	4	5	6	7-Very much so
How much did you behave within the experience as if the situation was real?							
How much was your emotional response in the experience as if the situation was real?							
How much were you thinking things like 'I know this isn't real' but then surprisingly finding yourself behaving as if it was real?							
To what extent were your physical responses within the experience (e.g. heart rate, blushing sweating, etc) the same as if it had been a real situation?							
How much did you behave as if Li was real?							
How much was your emotional response to Li as if she was real?							
How much were your thoughts in relation to Li as if she was real?							
How much were you thinking things like 'I know this person isn't real' but then surprisingly finding yourself behaving as if Li was real?							
How much did you have physical responses (such as change in heart rate, blushing sweating, etc) towards Li as if she was real?							

2. Please list as many objects in the room and items of clothing on the character as you can remember.

3. Please add below any comments you might have related to the most recent experience.

Demographic Questionnaire

1. Please insert your name:

2. Please insert your age:

3. Tick or write your gender:

Male	Female	Prefer not to say	Other:

4. How many times have you used Virtual Reality (VR) before?

0	1	2-10	11-50	50+

5. On average, in the past year, how many hours did you play games *per week*?

less than 1h	1h-5h	6h-10h	11h-30h	more than 30h

6. What type of video games do you play the most? Please order the following genres based on which ones you play the most:

Please order the following by drag and drop using the mouse

Strategy, Action/Shooter games, Platformers, Adventure, Esport games, Role-Playing, Simulation, Action/Fighting games, Puzzle games, Indie games

7. If you have any notes about the question above (order of games genre) please mention it here. If not, please add "-"

8. Please list the games you played the most in the last year.

9. What do you play videogames on primarily? Please select maximum 3 platforms:

Browser/Facebook (other social media) Gaming; Mobile (Phone/Tablet) Gaming; Games Consoles; PC; VR; Other:

10. Is this scenario/experience reminding you of any games you've already play/know of? If so, can you provide a few details?

11. Please add below any final comments you might have related to the whole experience.

I see myself as someone who...

	Disagree strongly	Disagree a little	Neither Agree nor Disagree	Agree a little	Agree strongly
is reserved					
is generally trusting					
tends to be lazy					
is relaxed, handles stress well					
has few artistic interests					
is outgoing, sociable					
tends to find fault with others					
does a thorough job					
gets nervous easily					
has an active imagination					

Onboarding questionnaire

- **Occupation:**
- **Average usage of mixed reality devices (e.g., HoloLens) in the past 6 months:**
[] Never used it; [] <1 per month; [] 1-3 times a month; [] 1 per week; [] 2+ times per week;
- **Please select the gender you most identify with:**
 - *[] Woman; [] Man; [] Prefer not to say; [] Non-binary/gender diverse; [] Self-Described;*
- **Please select your age group:**
 - *[] 18-25; [] 26-35; [] 36-45; [] 46- 55; [] 56-65; [] >66;*
- In this study you will be embodying two different styles of avatars. Have you ever embodied an avatar in VR or MR? If so, what was the experience like?
- In this study you will be embodying two different styles of avatars. Could you say in a few words **how do you feel** about having avatars represent you?
- In this study you and your colleagues will be embodying avatars in a work meeting. How do you think the **dynamics** of the meeting will be like?
- Are you nervous or hesitant about having work meetings in MR while being embodied by avatars?
- How would you feel about having the following types of meetings in MR while being embodied by avatars?
 - a meeting with your peers
 - a meeting with your manager/superior
 - a meeting with clients/customer
- What do you think are the **advantages** of having an avatar representation in MR meetings?
- What do you think are the **disadvantages** of having an avatar representation in MR meetings?

Daily post-meeting questionnaire

After each meeting, the participants will fill in the following questionnaire:

1. Please answer the questionnaire with the most recent meeting in mind:

- *I felt engaged in the meeting.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *I felt that MY COLLEAGUES were engaged in the meeting.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The avatars communicated like my colleagues.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The appearance of the avatars affected THE MEETING TASKS.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The appearance of the avatars affected HOW COMFORTABLE I FELT in the meeting.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The appearance of the avatars mattered to me*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *I felt that I was in the presence of my colleagues.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *I could identify my colleagues.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *I perceive my colleagues' avatars as being only computerized images, not real people.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *There were obvious unnatural nonverbal behaviours from my colleagues' avatars.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The avatars' nonverbal behaviour was APPROPRIATE for the context.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *The avatars' nonverbal behaviour was USEFUL for understanding my colleagues.*

<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>

- *Please include any comments you might have:*

2. Questionnaire on perceived mood, own mood and on the level of information each cue had on understanding the partner's mood.

If there are going to be groups of two, only the questions about COLLEAGUE1 will be asked (the questions about COLLEAGUE2 will be removed):

- Please rank the following items based on the level of information each gave you in understanding your colleague COLLEAGUE1 mood/state of mind (from most to least useful):
 1. C1's choice of words
 2. C1's tone of voice
 3. C1's movement/gesticulations
 4. C1's gaze
 5. C1's facial expressions
- Please rank the following items based on the level of information each gave you in understanding your colleague COLLEAGUE2 mood/state of mind (from most to least useful):
 1. C2's choice of words
 2. C2's tone of voice
 3. C2's movement/gesticulations
 4. C2's gaze
 5. C2's facial expressions
- Select how much you agree that your colleague COLLEAGUE1 had the following perceived moods:

	<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>
<i>optimistic</i>							
<i>focused</i>							
<i>annoyed</i>							
<i>stressed</i>							

- Select how much you agree that your colleague COLLEAGUE2 had the following perceived moods:

	<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>
<i>optimistic</i>							
<i>focused</i>							
<i>annoyed</i>							
<i>stressed</i>							

- Select how much you agree that YOU had the following moods:

	<i>Strongly disagree</i>	<i>Disagree</i>	<i>Somewhat disagree</i>	<i>Neither agree nor disagree</i>	<i>Somewhat agree</i>	<i>Agree</i>	<i>Strongly agree</i>
<i>self-optimistic</i>							
<i>self-focused</i>							
<i>self-annoyed</i>							
<i>self-stressed</i>							

B

Study Information Sheets & Ethics Forms

Below are the ethics documents and the participants' information forms for Chapter 4. The documents for Chapter 5 could not be made public as the documents were internal and the project handled employees' data.

We present the documents for the *Study analysing the participants' engagement with a Non-Player Agent (NPA) in Virtual Reality (VR) Environment* represents the project proposal document for the Ethics board. Following that, the *Postgraduate Research Ethics Declaration Form* shows the signed ethics declaration to run the study. Following that, there is the *Participant Information Sheet*, that volunteers received before signing the consent form. Finally, there is the *Informed consent form* that all volunteers signed before taking part in the study.

Georgiana Cristina Dobre

Student ID: 33346662

Study analysing the participants engagement with a Non-Player Agent (NPA) in Virtual Reality (VR) Environment

Attachment for Computing Department PGR Ethics Declaration Form

Q1 Human Participants

This study will look at detecting different ways people engage with Non-Player Agent (NPA) in social interactions. In this case, the NPA will be programmed to give a monologue about a past experience. The participant will be wearing a head-mounted display (HMD) and hold in each hand a controller. They will be able to move in the virtual space using a thumbstick (joystick) from the controller. Their physical movements are also mapped in the virtual space. The experiment is expected to take 45 minutes to one hour, including introduction, position synchronization, data recording, questionnaires and debriefing.

Experimental Conditions

Four experimental conditions will be administered in this specific order for each participant:

1. Interaction with no instructions (baseline)
2. Interaction with instructions to gain the Non-Player Agent's trust
3. Interaction with instructions to violate social norms
4. Interaction with instructions to explore the room and try to interact with objects in the room.

In the baseline condition, the participant will be put into the virtual space without being told how to act or whether to behave with an end goal in mind.

In the second condition, the user will be asked to engage in the same scenario such that at the end, they will gain the NPA's trust.

In the next condition the participant will be told to imagine that there is a camera in the scene that will record the overall interaction. They will be asked to engage in such a way that the interaction will be very entertaining for someone watching it on a video streaming platform such as YouTube or Twitch.

In the final condition the user will be asked to explore the room and try to interact with objects in the room.

During each condition the NPA will be giving a monologue about a past experience with both animation and audio. Each condition will take approximative 5 minutes and at the start of each, the participant will be asked to make different poses (such as T-pose) for synchronizing the head and hands data

Questionnaire

The participants will be asked to fill in short questionnaires about their experience. The questionnaires include questions about the overall impressions on the task, the monologue the NPA is providing,

participant's general feelings toward the character, regarding the participant's feeling of presence within the Virtual Environment [1] and about their personality using the Big Five Questionnaire [2] based on the Five-Factor Model (FFM). At the end of all four conditions, the participant will be asked to fill in a demographic questionnaire.

Participants

Participants will be recruited from among the working professionals within the 'Dream Reality Interactive' studio, the Goldsmiths students as well as other professionals who volunteer to take part in the experiment. Given the rich data captured and the application nature, a minimum 10-15 participants will be needed.

Risk factors

Participants engage in fairly trivial, passive activity in VR. They can move about using the controller thumbstick or physical movement. Issues around VR sickness are not expected, however, if these arise, participants are free to withdraw at any time.

Q2 Information relating to identifiable, living individuals

The experiment will take place in VR thus the participants will be using a Head Mounted Display (HMD) and hand controllers; the recorded data will consist of 6 Degree of Freedom (6Dof) information about the HMD and hand controllers. Thus, the participants will not be identifiable through these data. Apart from this, the experiment will be video recorded, and a screen capture recording made, to allow for analysis of any unexpected circumstances. this will not be directly used in the machine learning training.

Debriefing will be given. Personal data records will be stored separately and will not be published or made available outside the research team. Personal data, recordings and questionnaires will be destroyed after 5 years.

References:

[1] Slater, M., John MC., and Francesco M. "The influence of body movement on subjective presence in virtual environments." *Human Factors* 40.3 (1998): 469-477.

[2] Rammstedt, B. John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203-212.

**Department of Computing
Goldsmiths, University of London
Postgraduate Research Ethics Declaration Form**

Please complete this declaration if your research will involve the participation of anyone other than direct collaborators (i.e. supervisors, co-authors, etc.). Upon review, the PGR Committee may deem it necessary for the proposed research project to be reviewed by the College Ethics Review Board. You may not begin the proposed research until you receive confirmation from the PGR Committee.

Please return completed declarations to the **Postgraduate Research Tutor** or **Administrator**.

You are asked to answer the following questions, and give further details in an included document if you answer **YES** to *any question*.

1. Will the research or evaluation components of your project involve human participants, e.g. questionnaires, surveys, interviews or software testing?

YES / ~~NO~~

- a. If **YES**, will participants include vulnerable individuals such as minors, medical patients, drug users or elderly persons?

YES / ~~NO~~ / ~~NA~~

2. Will you report, in any form, information relating to identifiable, living individuals?

YES / ~~NO~~

3. If you are proposing a web-accessible application, will it store cookies on client browsers?

YES / ~~NO~~ / ~~NA~~

4. If any material you produce will include diagrams, figures, tables, illustrations, program code or other content not authored by you, do you have express or implied permission from the copyright holder, or know it to be covered by the College copyright license?

YES / ~~NO~~ / ~~NA~~

See notes overleaf

Student Name: Georgiana Cristina Dobre

Student ID Number: 33346662

Supervisor Name: Sylvia Pan; Matthew Yee-King

Is this research supported by an external funding body?: YES / ~~NO~~

If yes, which one: INNOVATE UK

Signed (Student): Georgiana Cristina Dobre

date: 09/07/19

Signed (Supervisor): 

date: 19/07/2019^[1]

Signed (PGRC chair or HoD): Aaron Gerow^[2]

date: 19/07/2019^[3]

Notes

1. If **YES**, briefly outline your study and confirm that you will:
 - a. Obtain written or electronic consent and tell participants that participation is voluntary, and that they may decline to answer particular questions;
 - b. Tell participants that their data will be treated in confidence, and they will not be identified in any project outputs without their consent.

If you plan to work with **vulnerable individuals**, you should bring this up with your supervisor urgently as you will need to seek approval from the College. Any work involving **minors** requires written consent of a parent or guardian and may require a DBS check.

2. It is expected that any data you collect will be anonymised and stored confidentially and securely so that individuals cannot be identified. If this will not be the case, explain the reasons to your supervisor before you collect such data. You should confirm that you will:
 - a. Obtain appropriate **consent** for use of collected information, confirm to participants that the data will be **destroyed / deleted** after a specified period, and allow them to request a copy of the information at any time after the study;
 - b. Explain to participants how the information will be used, and not use the data for any other purpose;
 - c. Ensure that data is secured from unauthorised access, physically and / or electronically. The College advises that data collected be stored on a Goldsmiths password-protected network drive or the One Drive.
3. If you use cookies, you should confirm that end-users will be given clear and comprehensive information about the content and purpose(s) of the cookies, and give them the option to refuse (which may comprise declining to access the application).
4. If you answer **YES** to question 4, you should ensure that:
 - a. All material not authored by you is appropriately acknowledged and cited according to academic, licensing and legal standards;
 - b. That you have express or implied permission to use it.

Study analysing the participants engagement with a Non-Player Agent (NPA) in Virtual Reality (VR) Environment

Participant Information Sheet

Thank you for thinking about taking part in this research study. This information sheet should tell you all you need to know before deciding whether to take part or not. If you have any further questions, please ask the researcher, who will be pleased to answer them.

Taking part in the study is entirely voluntary; there is no disadvantage to you if you decide not to.

Study Overview

The purpose of this study is to analyse how users engage with a Non-Player Agent (NPA) during a monologue-based scenario. If you take part, you will be presented with a VR experience played in four different conditions. You will be given different instructions for each condition. Each of them lasts for about five minutes. After the first condition, you will be asked to fill in a short questionnaire about the experience. Additionally, there is a short demographic questionnaire to be filled in at the end of these four conditions. Afterwards, in debriefing, you will have the chance to learn more about the project and you will be given the opportunity to give any impressions or feedback. The whole procedure will take 45 minutes to one hour.

The four conditions will be recorded with a video camera and by screen-capture. This will allow us to check back on any unexpected circumstances, including any points raised in debriefing.

Risks

There are no major risks associated with taking part in this study. Some people experience nausea in VR: if this does happen to you, close your eyes and remove the headset; you may wish to withdraw from the study in this case.

Data

Data from this study will be used for the creation of framework for an NPA behaviour, for the researcher's PhD thesis, and may also be reported in published journal papers or conference proceedings. Questionnaire responses will be linked through a unique ID number, so that this data can be analysed anonymously. Any publication will present only the analysis of this anonymised data. Your personal data and video recordings are confidential, and only the research team will have access to them. If any of your comments are quoted in any publication, this will be in an anonymous form that cannot be linked back to you.

Personal data will be stored securely on Goldsmiths' OneDrive. Anonymised questionnaire data and video and screen-capture recordings will be uploaded to Goldsmiths secure One Drive, and any local copies destroyed immediately. All data will be stored for 5 years.

Withdrawal

You are free to withdraw from the study at any time, without giving a reason. If you do, then your personal details and any questionnaire responses collected up to that point will be removed from the study and destroyed.

Research Team

Principle Researcher: Georgiana Cristina Dobre (c.dobre@gold.ac.uk)

Supervisory Team: Dr Sylvia Pan, Dr Marco Gillies, Dream Reality Interactive, Maze Theory

This research is funded by an Innovate UK grant for a project in collaboration between Goldsmiths University of London, Dream Reality Interactive and Maze Theory.



Participant Number	
---------------------------	--

Informed consent form

Informed consent for a study analysing the participants engagement with a Non-Player Agent (NPA) in Virtual Reality (VR) environment

Please tick the appropriate boxes **Yes** **No**

1. Taking part in the study

I have read and understood the study information dated I have been able to ask questions about the study and my questions have been answered to my satisfaction.

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

I understand that if I do decide to withdraw, anonymised data can no longer be removed from the study.

I consent voluntarily to the video recording of my participation and to the screen-capture recording of my interaction in the Virtual Reality Environment.

I understand that taking part in the study involves a series of short VR experiences, with brief questionnaires, my responses being recorded in digital form.

I understand that, in taking part in the study, there is a slight potential risk of VR induced nausea. If this becomes a problem, then I understand that I am free to stop immediately and withdraw from the study.

2. Use of the information in the study

I understand that the data collected during the experiment will be used in the creation of framework for an NPA behaviour, as part of the researcher’s PhD thesis, and may also be included in published journal papers or conference proceedings.

I understand that personal information collected about me that can identify me, such as my name or any video recordings made, will not be shared beyond the study team.

3. Signatures

Name of participant [IN CAPITALS] Signature Date

Name of researcher [IN CAPITALS]

Signature

Date

4. Study contact details for further information

Georgiana Cristina Dobre, email: c.dobre@gold.ac.uk

Sylvia Pan, email: x.pan@gold.ac.uk

Marco Gillies, email: m.gillies@gold.ac.uk

C

Monologue Script for Chapter 4

1 INT. BOUDOIR

Li is sat on the edge of the bed, staring directly at the player. She follows the players gaze up to the musical bird cage hanging from the ceiling.

LI

It is a thing of beauty, isn't it?

Li lowers her eyes and rests them back on the player.

Warmth monologue spoken with a warm, gentle tone.

LI (CONT'D)

When I was small, my mother would wind it up for me before bed. I'd listen to it's sweet melody until the twirling rose tinted finch cast a dreamy blanket over me.

Whenever I turn it on now, it's as if she's still here with me.

Do you know that feeling? When someone loves you unconditionally, and their warmth cradles you, even when they can't be with you

Li stares into the distance. Complentative.

Happy monologue spoken with a smile.

LI (CONT'D)

She would take me to the park, the one where tens of those finches would gather in the trees, tweeting.

Li lets a small laugh.

LI (CONT'D)

She'd wave her arms in despair when I'd run at the trees, scattering the little things.

Li stares directly at the player and a smile breaks out across her face.

LI (CONT'D)

But she never told me off. Instead, she'd pat the grass and we'd sit snuggled up waiting quietly for the birds to resetttle. They always did and then she'd say, "See, no harm done." And she was always right.

Their darting black eyes as they tweeted unknowable messages to each other amused us. We'd make up stories for them, where they came from, what their dinner plans were, and we'd laugh at our own ridiculousness.

Li shifts uncomfortably and her eyes narrow a moment as a thought crosses her mind. She goes back to staring off into the distance.

Wistful monologue spoken with a sombre tone.

LI (CONT'D)

That's the only place we could laugh freely. The park with the rose finches. They've built apartments on it now. No longer can I ever go there. I wonder what happened to all the finches? Maybe they found a new home.

Li stares directly at the player again, her brow slightly crumpled.

LI (CONT'D)

Do you think they would have found a new home?

Li shakes her head briefly and her shoulders slump over a little bit.

LI (CONT'D)

No, they're like me, still looking for somewhere else to call home. I often imagine them happy in some other park. But the image doesn't come naturally to my mind, so it can't be true. I feel sorry for the finches, don't you?

Li sighs and pulls her back up straight.

Sad monologue spoken with a tone that's about to crack into tears.

LI (CONT'D)

It's as if no one cared about them. I know that feeling. My father...

(pause)

... he didn't seem to care about us much. I used to think I could make him, if I loved him enough for both of us. But guess what? It didn't work. It didn't change a thing and....

Li 's voice cracks.

LI (CONT'D)

... and when she died, he still didn't seem to care then either. How can that be? How can a man not love such a beautiful woman.

Li wipes her eye and stands up.

Angry monologue spoken with an increasingly harsh tone.

LI (CONT'D)

I hated him the day he didn't come to the funeral. Would
it have hurt him to pay his respects?

No!

I asked him why afterwards and he said it was because of
him. My uncle. "What it's got to do with my uncle? I
shouted.

(Li's tone quietens for a moment)

I didn't understand at first. Of course my uncle loved my
mother, who wouldn't.

(A brief smile fleets across Li's face)

She had a special soul.

(Anger returns to Li's tone)

I didn't want to believe his words at first, it wasn't possible
to think that all those times me and her sat huddled
waiting for the finches, she was thinking of him, my uncle.
Dreaming of the day they would be together.
But it wasn't her fault! It was his!
My uncle destroyed my life, he took it away from me.

Li gestures to herself with her hands.

LI (CONT'D)

He's got me trapped here you know. He doesn't care
about me either. She's the only one who ever did and
she's gone!

(pause)

What are you staring at? Don't pretend to care about me.
I know you don't.
Get out!

(pause)

I said get out!

D

Additional Stats for Chapter 5

Order	Optimistic	Focused	Annoyed	Stressed
CR	0.239	0.168	0.867	0.658
RC	0.060	0.117	0.134	0.440

TABLE D.1: P-values from the paired two-tailed t-test on the self-reported emotional state rating while embodying C or R avatars in order CR and RC (Cw1 vs Rw2; Rw1 vs Cw2)

Order	Week	Avatar	Optimistic	Focused	Annoyed	Stressed
CR	W1	C	0.584	0.975	0.638	0.170
RC	W1	R	0.034	0.337	0.071	0.023
CR	W2	C	0.274	0.102	0.701	0.629
RC	W2	R	0.999	0.223	0.430	0.556

TABLE D.2: P-values from regressions on the self-reported emotional states rating over time

Order	Week	Avatar	Optimistic	Focused	Annoyed	Stressed
CR	W1	C	0.829	0.550	0.776	0.626
RC	W1	R	0.053	0.916	0.024	0.133
CR	W2	C	0.477	0.666	0.250	0.569
RC	W2	R	0.175	0.000	0.014	0.418

TABLE D.3: P-values from regressions on the mapped error of the perceived emotional state over time