Faculty of Science

Department of Information and Computing Science

Human Computer Interaction

Master's Thesis

Role-Playing with Virtual Actors in an Online Communication Training

The Influence of Different Character Representations on User Experience and Learning Outcome

Author:

Lisa Wagensveld

Supervisors:

dr. C. van Nimwegen
Utrecht University

dr. ing. S.C.J. Bakkes
Utrecht University

B. van der Velde
Faculty of Skills

September 27, 2022

Abstract

Effective communication is an important workplace requirement and can be improved through soft skill training programmes. A frequently used method for training interpersonal skills is through role-play. Faculty of Skills, the training agency this research is conducted for, uses video role-play in their online training programmes. These videos are currently created by recording actors in a studio, however, this is an expensive and time-consuming process. Therefore, this research investigates whether digital alternatives (i.e., deepfakes and virtual humans) could replace actors in these videos without coming at the expense of the learning outcome and user experience. Thus, the following research question was formulated: *"How do different character representations in a virtual communication training influence the experience and learning outcome of trainees?"*.

To answer this, a between-subjects study was designed, in which participants (N = 103) were randomly assigned to either the A) actor, B) deepfake, or C) virtual human version. Participants completed a short training programme and filled out a survey afterwards. The results show that there is no significant difference in learning outcome between the three groups.

Additionally, participants did not rate the quality of the training significantly different. The quality of the characters, however, was rated significantly lower by participants from the virtual human group. Yet this version was rated as more inventive than the other two. No significant differences where found for other components of the user experience, such as how enjoyable, interesting, or attractive the training was. Furthermore, no significant differences were found in the feeling of engagement, presence or immersion.

Finally, four video characters (2 male, 2 female) of each version were rated based on several statements. The virtual humans were generally rated worse than the other two groups. An explanation for this could be due to the less natural sounding text-to-speech (TTS) voice, or perhaps due to factors such as the length of the video or design choices regarding the appearance of the character. On the other hand, gender of the character did not seem to influence the perception.

Based on the results, Faculty of Skills could potentially start tests with actual clients, in which they offer them personalised role-plays that are delivered faster and at a lower price. This way, it can also be verified whether these results hold up after a full-length training programme as well.

Keywords: *virtual humans, deepfakes, communication training, video role-play, user experience, learning outcome*

Contents

1 Introduction

Companies spend large amounts of money on training; specifically on training programmes that focus on the development of interpersonal skills (Schmid Mast et al., 2018; Riggio & Lee, 2007). Effective communication is an important requirement for professionals from many fields (Gartmeier et al., 2015). In order to teach professionals these skills, simulated scenarios need to be provided that allow them to practice the subtle nuances and complexities that are involved in communicating a message effectively (Chetty & White, 2019).

A frequently used method for training interpersonal skills is through role-playing (Schmid Mast et al., 2018). Typically in role-play, two or more participants act out a certain role in a predefined scenario to experience specific social situations, while being observed by other people. Next, their experience is reflected upon and they receive feedback on their performance from trainers or peers, or by watching a video recording of themselves (Sogunro, 2004; Nestel & Tierney, 2007; Salas et al., 2009). Role-playing is an experience-based form of learning (i.e., learning by doing), which is more effective in communication training than simply learning by instruction, according to Aspegren (1999).

Role-playing can also be done online, for example in a virtual environment. In a virtual environment the role-play partner is often a virtual human, which in short, is a 3D representation of a human (Schmid Mast et al., 2018). They can speak with a (prerecorded) human voice or via text-to-speech (TTS), while moving their lips in synchrony with what they are saying. Virtual humans can be agents (i.e., entirely preprogrammed) or avatars (i.e., a representation of a real human in the virtual world, controlled by this real human) (Fox et al., 2015). An advantage of using an agent as a conversation partner is that no real human is needed to control them, as they are programmed to react in a certain way. This is a very cost efficient way to train interpersonal skills, because the training partner is replaced with a computer algorithm (Schmid Mast et al., 2018).

1.1 Research problem

This research is conducted for Faculty of Skills (FoS), a training agency based in Utrecht. They offer online training programmes for developing a variety of communication skills to improve workplace communication. FoS has developed a tool called TrainTool, which is an online learning environment in which their clients can practice their skills through video role-play (Faculty of Skills, 2021a). In these role-plays, trainees are presented with a predefined problem. Next, they watch a pre-recorded video of their role-playing partner, which they then have to respond to. Faculty of Skills wants to offer their customers the best learning experience for the best price, so the production of the video content (i.e., the costs and time) plays a large role in this.

FoS currently records their role-play videos in a studio, using real-life actors, which is a costly and time-consuming process. Therefore, this research will investigate new methods for creating these videos. To do so, an experiment will be created in which human actors are compared to alternative characters. For the experiment, a segment of an existing training will be recreated, replacing the actors with different characters. The first alternative will be created using virtual humans, which are computer-generated characters that are designed to look and behave like real people (Swartout et al., 2013). In the second alternative the agent will be a deepfake; a hyper-realistic video that is created using artificial intelligence (AI) (Chawla, 2019). These alternatives will be compared to the original method of using actors to investigate whether they could be viable alternatives that do not come at the expense of the experience and learning outcome of the users.

Therefore, the following research question has been formulated: *How do different character representations in a virtual communication training influence the experience and learning outcome of trainees?*

The next section goes into more detail about Faculty of Skills and their training programmes. Furthermore, the literature section will elaborate on the difference between virtual humans and deepfakes, and how they have previously been used in learning environments. Then, the methodology section will explain more about the experiment and how the video content was recreated. Finally, the results from the experiment will be analysed, discussed, and concluded.

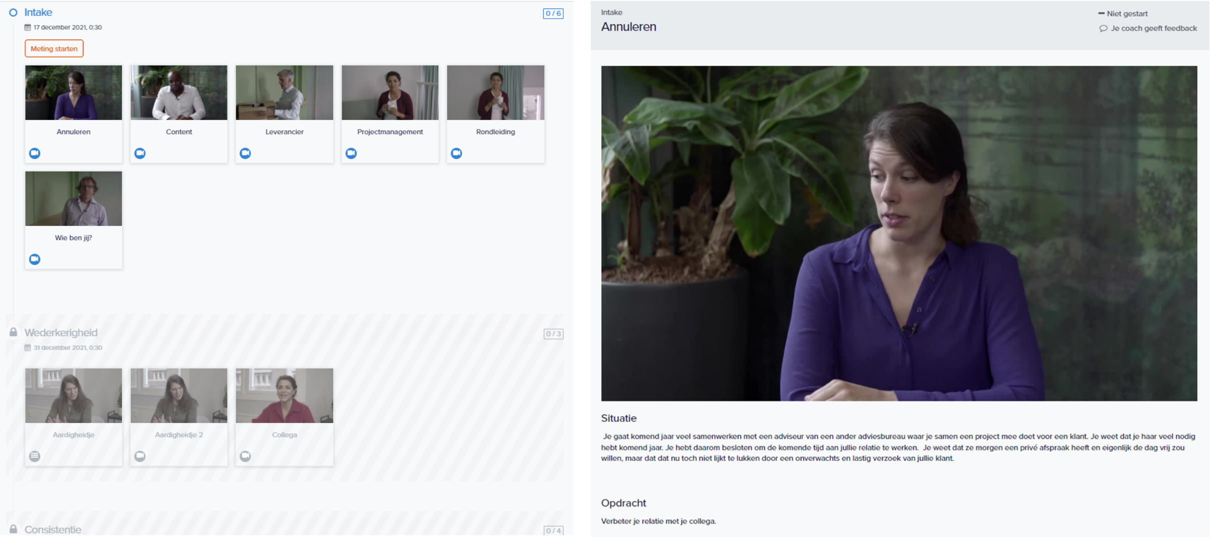
2 Case study

This section will explain more about Faculty of Skills, TrainTool, and their training programmes. FoS offers online training programmes for developing communication skills. Their goal is to help their clients, who range from students to ambitious professionals, communicate better in their workplace. This allows their clients to have more impact, reach their goals faster, and work well together with others (Faculty of Skills, 2021a).

To achieve this, Faculty of Skills has created TrainTool, an online tool that can be used to practice skills in online video role-plays. Clients can increase their knowledge about communication skills and practice them directly in practical situations. Afterwards, they receive personal feedback from either a coach or from Alix, the FoS AI coach.

Tens of thousands of employees of large organisations and students from the Netherlands and abroad participate in the online TrainTool programmes. Clients appreciate the ease of use, service, and solid software, especially in the field of privacy and security. TrainTool already has many features, but FoS keeps innovating and further developing their tool together with their clients and experts (Faculty of Skills, 2021c).

Faculty of Skills has developed a great variety of training programmes for the most sought-after skills. These can be used for personal training, or can be done with an entire team. Some examples of training programmes they offer are Influence according to Cialdini, Effective communication, Delivering bad news, and Applying for jobs. Besides their standard programmes, they also offer tailor-made programmes (Faculty of Skills, 2021b).



Situatie

Je gaat komend jaar veel samenwerken met een adviseur van een ander adviesbureau waar je samen een project mee doet voor een klant. Je weet dat je haar veel nodig hebt komend jaar. Je hebt daarom besloten om de komende tijd aan jullie relatie te werken. Je weet dat ze morgen een privé afspraak heeft en eigenlijk de dag vrij zou willen, maar dat dat nu toch niet lijkt te lukken door een onverwachts en lastig verzoek van jullie klant.

Opdracht

Verbeter je relatie met je collega.

Figure 1: Example of a TrainTool training and exercise.

Each training starts with an introduction and a practice role-play to set-up the webcam and microphone. Next, an intake is done to get a baseline measurement of the user's skills. They are presented with a situation which they have to respond to without any practice. Afterwards, the user will get to practice the skills in different role-play scenarios. They have to apply the theory they have learned and record themselves responding to the presented situations. Once all components are completed, every training ends with an outtake to get a final measurement of the user's skills. This way, the user's learning outcome can be expressed in numbers. An example of a TrainTool training is displayed in Figure 1.

To create each new training programme, many new role-play videos need to recorded. Creating these videos is a time-consuming and costly process, because these videos are currently filmed in a studio using actors. There are both internal and external costs involved in this process, but especially the latter are expensive. For example, they need to hire the studio and actors for several hours, on top of which they have to pay the cameraman for filming and editing the videos. Internal costs, on the other hand, are things such as the design sessions and developing the content.

The total production costs can be reduced by eliminating the external costs. Replacing the actors by virtual characters (e.g., deepfakes or virtual humans) could potentially make the video production process more time and cost efficient. Another advantage is that most virtual characters can talk via text-to-speech (TTS), which makes it easier to record the same programme in different languages.

Currently, the role-play videos in TrainTool are prerecorded. If virtual humans or deepfakes are perceived just as well as real actors and no difference in learning outcome is found, this could present more opportunities for interactive role-play (e.g., with more branching options) in the future. By using natural language processing (NLP), the virtual characters could respond to trainees in real-time. Therefore, the response of a trainee could result or 'branch' into different paths, each with a unique reaction to the trainee.

The next section will explain more about virtual humans, the contexts they are used in, and deepfake technology.

3 Literature

This section will explain more about what virtual humans are and how they have previously been used in virtual learning environments, for example in role-play exercises. Additionally, this section will discuss how the embodiment (i.e., the appearance) of virtual humans and the emotions they display can influence the way they are perceived. Moreover, the uncanny valley will be discussed, as this phenomenon is important to keep in mind when designing virtual characters. Finally, deepfake technology will be explained, because this technology was used to create the deepfake avatars for the training videos.

3.1 Virtual humans in learning environments

Virtual humans are anthropomorphic (i.e., human-like) animated characters that are incorporated into virtual worlds (Fox et al., 2013) and digital learning and training environments (Craig & Schroeder, 2018; Tan et al., 2020). Depending on the role virtual humans play in their environment, literature has used different names to refer to them (e.g., animated agents, conversational agents, pedagogical agents, embodied virtual agents, or synthetic humans) (Haake & Gulz, 2009; Zipp & Craig, 2019; Tan et al., 2020).

Within a learning environment, virtual humans are often referred to as *pedagogical agents* or simply *agents* (Craig & Schroeder, 2017, 2018). These agents are virtual characters that are controlled by a computer system. They provide information to users or interact with them in order to facilitate learning (Craig & Schroeder, 2018) and have shown to increase the user's time interacting with the environment (Lane et al., 2011) and to improve learning (Schroeder et al., 2013; Craig et al., 2015).

Pedagogical agents can humanize and enrich the interaction between humans and computers through facial and emotional expression, vocal speech, and body gesture (Tan et al., 2020). Because these agents often play the role of an instructor or peer (Kim & Baylor, 2006; Haake & Gulz, 2009), they can be used to replace human role-players. This allows people to learn from their interaction with the agent, rather than the agent simply assisting them with their problems. Within this context, the virtual human and the learner both play a predefined social role, so that the learner can achieve certain goals by using specific communicative skills (Lane et al., 2013).

3.2 Virtual humans in role-play

Hays et al. (2012) researched whether virtual humans can teach interpersonal skills to Navy officers. In their experiment, the class participated in a semi-structured role-play exercise between a student and a life-sized virtual subordinate. A turn-based branching narrative was used for the interaction, which means the learners chose to say one out of three pre-scripted responses after the subordinate's response. Their experiment was designed to investigate how well the virtual human functioned as a role-player, and to do so, they compared it to a role-play with a live human in the same scenario. A survey was used to evaluate how participants responded to the role-players, what their perception of realism of the virtual human and the social characteristics of the interaction was, and what their experience as a whole was like. Additionally, they collected physiological data (i.e., heart rate and perspiration) during the sessions. Participants were randomly assigned to either meet the virtual or live role-player and followed the same branching narrative. Hays et al. (2012) found that interaction with the virtual human was approximately as engaging as interaction with a live human in the same room.

3.3 The embodiment of virtual humans

Research has shown that how agents are perceived and how effectively they facilitate learning can be influenced by their design (i.e., their voice, speech patterns, and physical appearance) (Ozogul et al., 2013; Schroeder et al., 2017), which is why they should be mindfully designed.

Aspects of the embodiment of these agents, such as their physical appearance, form people's first impression of the agent, which influences the way humans interact with them (Baylor, 2005). Face-to-face communication is affected by various non-verbal variables, which are cues that are either behavioural or non-behavioural in nature. Non-behavioural cues exist of demographic variables (e.g., ethnicity, age, gender) and physical appearance variables (e.g., clothing, attractiveness of body and face) (Khan & De Angeli, 2007). Physical appearance can be subject to change, which is why for example, hair and eye colour, attire, and make-up can affect social reactions to an agent (Cowell & Stanney, 2005).

Previous research also found the importance of demographics in the embodiment of virtual agents. Cowell & Stanney (2005) discovered that users prefer interacting with agents that either look young, or match their own ethnicity. Moreover, Khan & De Angeli (2007) compared existing virtual humans based on ethnicity, age, and gender. They found that the majority of agents are portrayed as young adults. One reason for this may be that most users are young adults, who may prefer to interact with agents of a similar age group (Cowell & Stanney, 2005). Nowadays, however, users can range from young children to pensioners, which is why these age groups should also be considered when designing virtual agents (Khan & De Angeli, 2007).

Findings from a study by Hone (2006) suggest that the agent's gender could also have an influence on the user. This research found that embodied agents improve frustration reduction. More specifically, the study suggested that female agents are more effective than male agents in reducing frustration.

3.4 Virtual humans and emotions

Communication happens both vocally and through face and body movements, which is why besides the semantic content of a message, a lot of other information is conveyed, such as emotions, mood, and affective ties. This is classified as socio-emotional content (Ellsworth & Ludwig, 1972, as cited in Castillo et al., 2018). In general, machines do not process or produce socio-emotional information, however, people are still inclined to base their interaction with machines on interpersonal behaviour cues that come from the machine (Lee & Lee, 2007).

Bosse et al. (2016) designed an intelligent training system for public domain professionals (e.g., train conductors) to practice their aggression de-escalation skills. The virtual reality environment focused on specific situations with high realism and detailed interactions with the (photo-realistic) virtual agents to immerse the player in the game. Participants were randomly assigned to either have an aggressive virtual agent or an aggressive human as their conversation partner. Bosse et al. (2016) compared the impact of the conversation partner by measuring the physiological and subjective emotional state of participants before and after an aggressive outburst of the partner. This outburst was caused by the virtual or human partner suddenly getting extremely angry towards the participant and shouting at them. The results showed that both the virtual and human partner induced a substantial stress response, but that the impact of the human aggression was stronger than that of the virtual aggression. While the human response was stronger, the results still indicate that virtual humans can trigger responses similar to the ones caused in reality. An improvement that was identified, however, was related to the emotional aspect of the system. Some participants noted that their perceived sense of presence was limited because they did not actually feel the emotion of the virtual partner. The authors suggest combining the scenarios with haptic feedback or using a head-mounted display (as opposed to a flat screen) to create a more engaging and perhaps a more effective training tool. Overall though, considering user satisfaction, participants were mostly positive about the content of the virtual scenarios and the interaction with the characters.

3.5 The Uncanny Valley

In 1970, robotics professor Mori first defined the Uncanny Valley theory (Mori et al., 2012), which since then has played a prominent role in research on how humans react towards virtual agents and avatars (Yuan et al., 2019). According to this theory, people have more affinity for agents that are more realistic. This affinity increases as the agent becomes more realistic, however, when the agent becomes semi-realistic the affinity point drops significantly, as shown in Figure 2. This drop occurs because a partially realist agent triggers a certain unease in humans (Mori et al., 2012). However, once the human likeness gets more real, it can subconsciously be accepted Fang et al. (2021).

There are many different theories that try to explain how the Uncanny Valley is caused. One theory argues that the drop in affinity is caused due to perceptual surprise (Mitchell et al., 2011). According to this theory, the first 100 to 300 milliseconds after seeing what could potentially be a face, the subconscious originally concludes that the almost-human character is a human. Therefore, it creates an expectation of its humanity. Then, it directs the conscious attention to focus on the character, however, when it determines that the character is not an actual human, this surprises the conscious attention. Finally, this surprise triggers a negative emotion.

Another theory argues that almost-human characters are perceived to be human, but because they have features that are less than perfect, they are often dehumanized (Wang et al., 2015). This dehumanization is a process in which a human is perceived to lack the attributes that encompass what it means to be a human. For example, this also occurs when seeing a person from an out-group that is different from the in-group of similar people. In the case of machines, dehumanization triggers negative emotions, because machines lack emotions.

It seems like the drop in affinity for the character is not deliberate, but driven by subconscious processes instead. The same goes for the final theory that will be discussed. Mori et al. (2012) argues that responses to almost-human characters are subconscious reactions for self-preservation. Almost-human characters are perceived to be humans expressing a psychopathic personality disorder (Tinwell et al., 2011). Because they fail to show emotions and behave in the same way as healthy humans, these characters are perceived to be dishonest and insensitive.

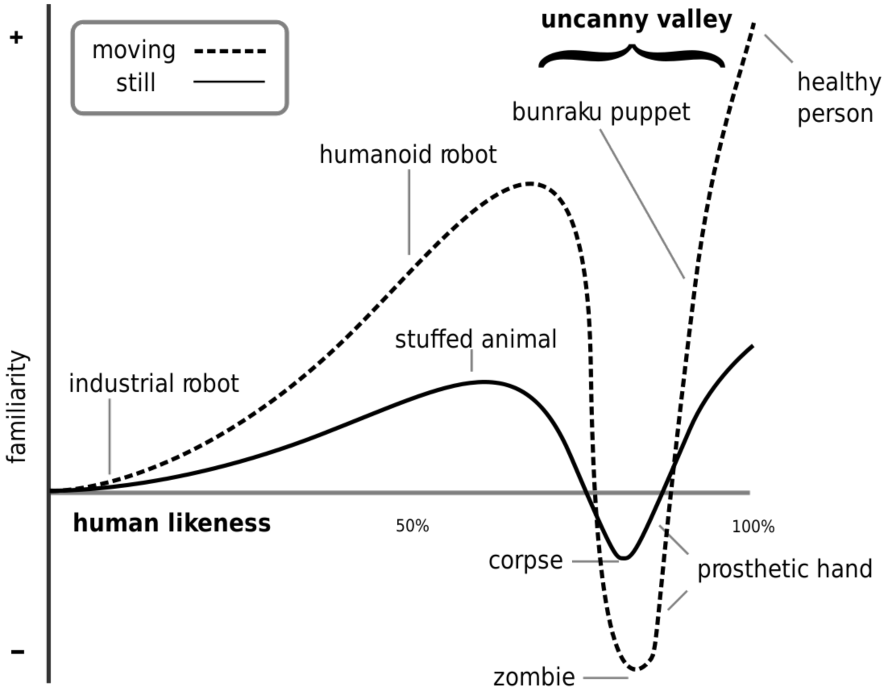


Figure 2: The Uncanny Valley.

Virtual humans aim towards the ‘healthy person’ point on the Uncanny Valley graph, as the goal is to simulate lifelike human behaviour and appearance. In order for their behaviour to be considered both verbally and non-verbally convincing, virtual humans should possess a high degree of autonomy, be proactive, and display emotional artificial intelligence (Draude, 2011). According to McDonnell et al. (2012), virtual characters receive negative reactions when the character has human texture maps and realistic human skin patterns, but does not have realistic eyes and skin shaders.

3.6 Deepfakes

Because of recent technological advancements it has become easier to create deepfakes. Deepfakes are hyper-realistic videos using face swaps that usually barely leave trace of manipulation (Chawla, 2019). They are made by artificial intelligence (AI) applications that merge, combine, replace, and superimpose images and video clips to create fake videos that appear to be real (Maras & Alexandrou, 2019). Deepfake technology can generate, for example, a humorous, pornographic, or political video of a person saying anything, without the consent of the person whose image and voice is involved (Fletcher, 2018; Day, 2019). Juefei-Xu et al. (2021) define a deepfake as "the creation and the manipulation of facial appearance (attributes, identity, expression) through deep generative approaches". They classify deepfakes in the following four categories: 1) entire face synthesis, 2) attribute manipulation, 3) identity swap, and 4) expression swap.

When deepfake technology is successfully implemented, it can be applied to a variety of use cases in which it is hard to distinguish from real human communication (Boneh et al., 2020). It can be used both maliciously (e.g., spreading fake news) and benignly (Juefei-Xu et al., 2021). Responsible use of this technology can provide clear benefits to society (Boneh et al., 2020). A benign, positive example of this is Synthesia, which uses deepfake technology to allow companies to create cost-effective synthetic training videos.

Now that there is a better understanding of what virtual humans and deepfakes are, the methods section will go into more detail about how these characters will be created for the experiment.

4 Methods

This section will elaborate upon the experiment and all the materials that were used to create it. For the experiment, participants were instructed to complete a training in TrainTool and to fill out a survey afterwards. They were divided into three groups, each exposed to a different type of video character (i.e., A. actors, B. deepfakes, or C. virtual humans) to role-play with during the training. Before explaining the experiment and its procedure in more detail, the research questions and hypotheses will be stated. Next, the recruitment of participants and some ethical considerations are discussed. Then, an overview of the materials will be given, including the chosen training programme and all the materials that were used to recreate the original actor videos into the deepfake and virtual human versions. This subsection will also elaborate upon how TrainTool was used to measure the learning outcome and how user experience was measured through the survey. Finally, the procedure of the experiment and its variables will be discussed.

4.1 Research questions and hypotheses

To answer the main research question, *"How do different character representations in a virtual communication training influence the experience and learning outcome of trainees?"*, the following null hypotheses have been formulated:

- H0: There is no significant difference in learning outcome between different character types.
- H0: There is no significant difference in experience between different character types.

To gain more insight into the preferences of participants, some additional research questions and corresponding null hypotheses have been formulated:

- RQ1: How does the character's gender influence the way it is perceived?
  H0: The character's gender does not influence the way it is perceived.
- RQ2: How does the character's voice influence the way it is perceived?
  H0: The character's voice does not influence the way it is perceived.

4.2 Participants

TrainTool is used by students, young professionals, and more experienced professionals over the age of 50. The average age of TrainTool users is estimated to be between 30 and 35 years old. Therefore, anyone that studies or works is eligible to participate in the experiment. Convenience sampling was used to recruit most of the participants (e.g., friends and family members). Furthermore, snowball sampling was used to recruit friends, family members and colleagues of people that already participated in the experiment. This made it easier to recruit people older than 30. Participation was rewarded with a €5,- gift card from BOL.com.

In total, 103 people participated in the experiment, of which 33 in version A (actor), 36 in version B (deepfake), and 34 in version C (virtual human). The age of participants ranged from 18 to 69 (M = 26.84, SD = 10,24). Figure 3 gives an overview of the participants' gender and employment status per version.

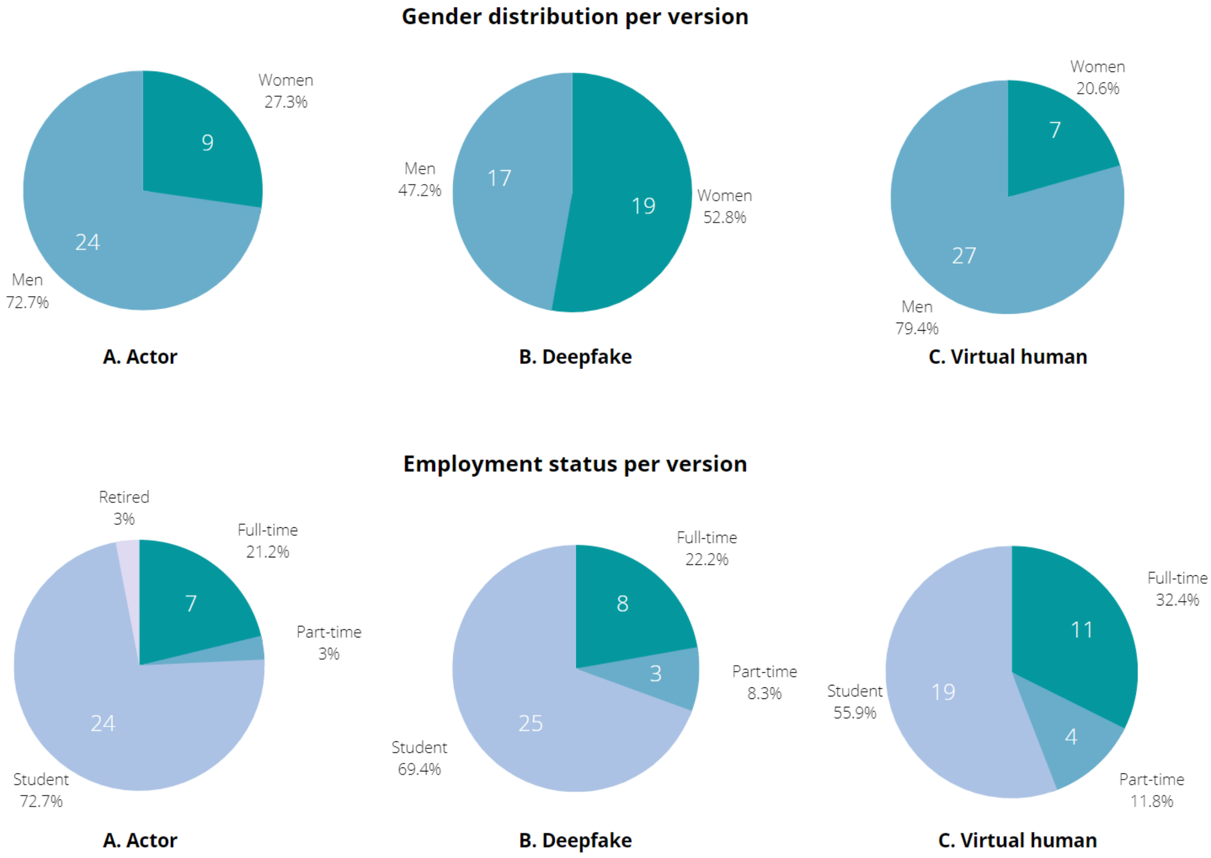


Figure 3: The distribution of gender and employment status per version.

4.3 Ethical considerations

When doing a training in TrainTool for the first time, it can be awkward for trainees to record themselves talking. They often get used to this after a while, but in this experimental setting they might not have enough time for this. Because these recordings are seen as sensitive data, participants were informed about how their data was used and stored through an informed consent form.
The following (mandatory) data was collected upfront:

- name
- e-mail address
- personal password (randomly generated, one-way encryption)

The following data was generated during usage:

- name of the training(s) they are enrolled in
- audio or video recordings that user decides to save and share (generated while following the course)
- information about participation rate, progress, and scores such as answers given, time spent, and assessment scores

When the participant's TrainTool license expired (i.e., after the research was finished), their answers and recordings were automatically hidden from themselves and potential coaches. Moreover, all user data, including their account and video recordings, were deleted 45 days after their license expired (Faculty of Skills, 2022). Participants were informed about this via the informed consent form found in Appendix A, which they had to read and sign before participating in the experiment. Participants were also told they could withdraw at any moment and request their account to be deleted if they changed their mind about participating. However, the form also stated that they would only be rewarded for their participation if they had finished the entire experiment (i.e., the training and survey).

4.4 Materials

4.4.1 Devices

Participants were instructed to do the training on either a laptop or desktop with a webcam and microphone. They were able to access the training via their browser (Chrome, Firefox, or Edge). While there is also a TrainTool app available for phones and tablets, the web version is used more than the mobile version. By only using the web version of TrainTool, screen size was removed as a variable that could have influenced the perception of the characters.

4.4.2 Choosing the training

Faculty of Skills offers a great variety of training programmes. Some of them teach more general skills (e.g., Influence according to Cialdini, Effective communication), while others are more context specific (e.g., Dealing with aggression in the medical world). Additionally, some programmes involve more emotion, such as dealing with agression or delivering bad news. Research found that virtual humans displaying aggressive behaviour can trigger stress responses in participants, however, the impact of human aggression was found to be stronger (Bosse et al., 2016). Even though this would have been interesting to investigate further, a more general training fit better within the scope of this research. This way, there were fewer variables that could have influenced the results.

In consultation with a FoS learning designer the 'Giving feedback' training was chosen for the experiment. This Dutch training teaches how to give positive and constructive feedback, how to apply the 4G-model, how to express feedback respectfully, and how to respond to feedback received from others. The 4G-model consists of the following components: Gedrag (behaviour), Gevolg (consequence), Gevoel (feeling), and Gewenst gedrag (desired behaviour). In short, the observed behaviour is described in a neutral tone, followed by the consequences of this behaviour. Next, the person indicates how this behaviour makes them feel, and finally they tell the other person what kind of behaviour they would like to see in the future.

The training has been duplicated to a research environment of TrainTool. In this environment, the FoS learning designer adjusted the training to make it more concise. The actual training programmes usually take multiple hours (spread over multiple days or weeks), which would have made it very difficult to find a sufficient amount of participants for the experiment. Therefore, it was important that the training did not take too long, yet also contained enough content to be representative of a real TrainTool training. It was decided that the experiment should take no more than 30 minutes for participants to complete. The training itself consisted of an intake, three practice rounds and an outtake. This way, the difference between the baseline and final measurements (i.e., the learning outcome) could be measured.

In total, the adjusted feedback training consisted of 16 prerecorded videos. The training used the same video for the intake and outtake, so it contained 15 original videos. All together, there was 290 seconds worth of video content, which is a little under five minutes. Participants had to record themselves six times. This included setting up their webcam during the introduction role-play.

4.4.3 Choosing the actors

Faculty of Skills choose actors that fit the context of the programme. For instance, they pay attention to details such as the degree of professionalism and age. Because an existing training had been chosen for the experiment, the actors had already been thoughtfully selected. Therefore, an attempt was made to create virtual humans and deepfakes that resembled the original actors as closely as possible.

4.4.4 Creating videos with virtual humans

Several tools for making virtual humans were explored, and two of those could potentially be used to create the training videos for the experiment. The first option that was explored was the Digital Human Creator by UneeQ. These Digital Humans are mainly deployed as chatbots, for real-time interaction with the user. They draw on a natural language processing (NLP) platform, which is essentially the IQ of the interaction. Additionally, by using the emotionally driven communication methods of speech and body language, UneeQ adds EQ to their Digital Humans. This way, Digital Humans can listen, understand and speak with users just like actual humans do. Their physical appearance looks realistic, however, the pricing started at $4.000 (USD) per month, which meant that this tool was not a viable option for the experiment.

The second tool that was found was the MetaHuman Creator by Unreal Engine (UE). This creator allows users to quickly create high-fidelity digital humans. There is a large variety of MetaHuman Presets (i.e., premade characters) readily available. Some of these are portrayed in Figure 4. They have different age, gender, and ethnic features, which can be adjusted manually or by blending several Presets together. The MetaHuman Creator is a free cloud-streamed app, which allows the user to create digital humans in their browser. When the human is finished, it can be downloaded and exported to Unreal Engine using Quixel Bridge. Both of these programmes can be downloaded for free.

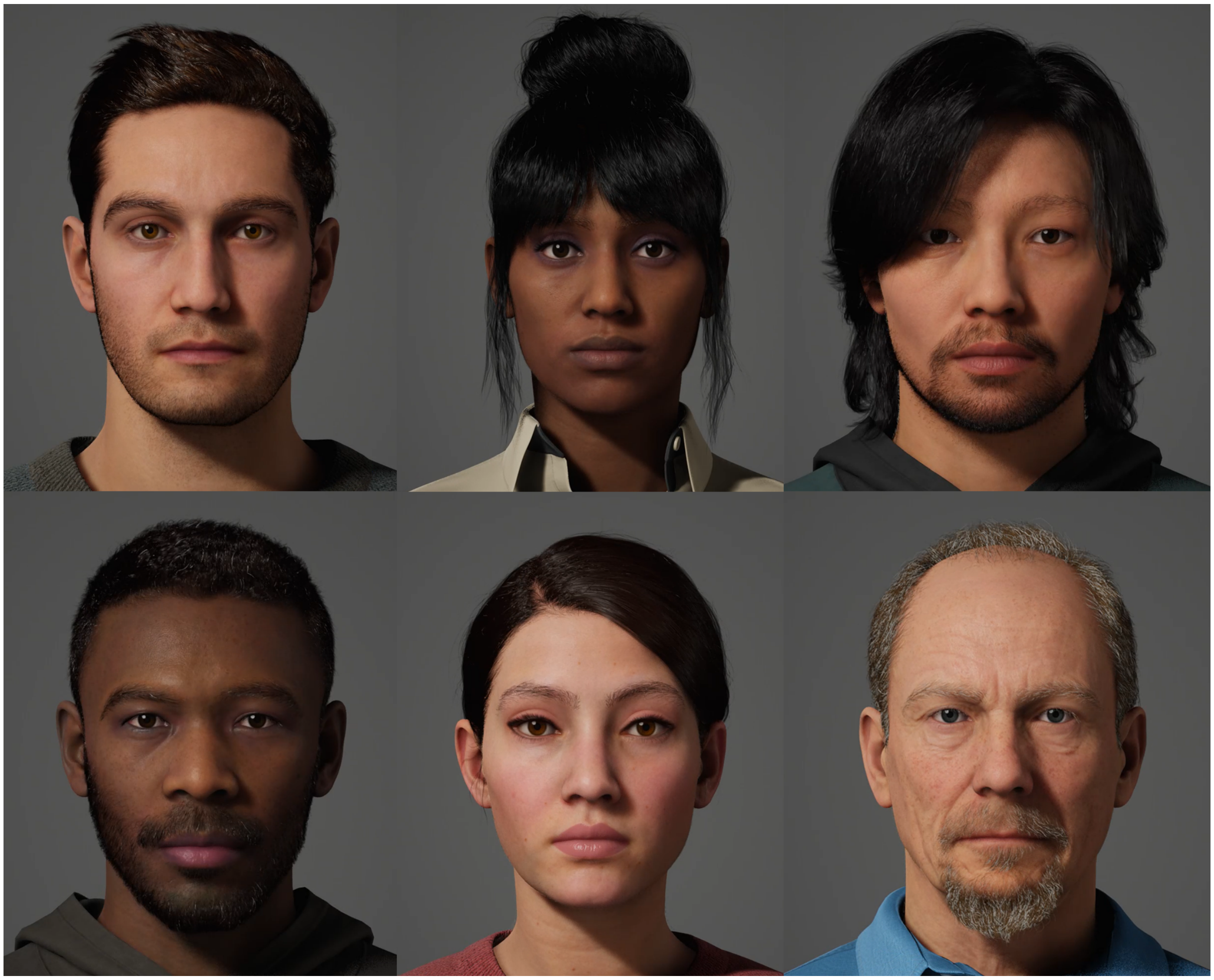


Figure 4: A variety of MetaHuman Presets.

While UneeQ's Digital Human is specifically designed to function as a chatbot to have real-time interaction with, Unreal Engine's MetaHumans can be used as characters in for example games and movies made in UE. Unlike UneeQ's Digital Humans, however, MetaHumans do not have a built-in AI voice. Therefore, some additional tools were required to make a training video with the MetaHumans. First, their face needed to be able to move while they were speaking. This facial animation was done through face tracking, allowing the MetaHuman to mimic the facial movements of a real person speaking. Unreal

Engine has made an app for this, called Live Link Face App, which is available on iOS (13 and up).

Next, the MetaHuman needed a voice. For this, VoiceMaker was used, which is a web-based TTS (Text-to-Speech) tool. It allows its users to convert their text to speech in human-like speaking voices. Moreover, the tool offers customizable audio style, voice speed, pitch, and volume. Pauses and emphasis can be added as well. They currently have over 600 AI voices in more than 70 languages available, including Dutch. A premium plan costs $10 per month. This plan allows more text characters and provides more voices. Instead of only standard TTS, this plan also provides Neural TTS, which produces more natural and human-like voices.

In conclusion, the following devices, programmes, and apps were used to create the videos with virtual humans:

- Laptop or desktop (for recommended specifications see UE website)
- iPhone (or iPad, iOS 13+ and TrueDepth camera required)
- MetaHuman Creator (browser)
- Unreal Engine 4.27
- Quixel Bridge
- Live Link Face App
- VoiceMaker (browser)

4.4.5 Creating videos with deepfakes

The deepfakes were created using Synthesia. Synthesia offers a large variety of AI avatars. These avatars are all based on video footage of real actors, two of which can be seen in Figure 5. Then, the footage is taken through Synthesia's AI system to create new videos from text input. They offer a large selection of synthetic voices available in 60+ languages, including Dutch, ready to be used with all the avatars. Additionally, one can also upload their own audio instead of using using TTS. Furthermore, backgrounds can be customised and text can be added to the videos. Synthesia's pricing starts at $30 per month for 10 video credit (= 1 credit is 1 minute of video) and they also offer corporate plans. Faculty of Skills already had an account available to create the videos with.

The first few videos were created using an AI voice, but even after tweaking it in Voicemaker it still sounded quite unnatural. Compared to English AI voices, there are only a limited amount of Dutch voices to choose from and these are of lower quality as well. Therefore, the videos ended up being made with the original audio files from the actor videos. This sounded a lot more natural, as the contrast between a real-looking person and a real voice was a lot smaller than when it spoke in an AI voice.

Figure 5: Examples of Synthesia avatars.

4.4.6 Measuring learning outcome in TrainTool

Each TrainTool training consists of an intake (baseline measurement), the actual training content and exercises, and an outtake (final measurement). The baseline measurement was done before participants had practiced. Participants were presented with a brief situation to which they had to respond. They got two tries to record their answer, but in case they used their second try this always counted as their final answer. Then, the recording was assessed on six criteria (e.g., statements such as "I give an observation of the behaviour") that could either be correct or incorrect. After completing the training, participants had to respond to a similar situation and apply the skills they had learned during the training. This recording was assessed in the same way.

In TrainTool, there are two ways to assess those criteria: through self-assessment (i.e., participants assess themselves) or through coach assessment (i.e., a FoS coach assesses participants). Experience showed that participants are perfectly capable of assessing themselves and are often even stricter than the coaches. Additionally, this option was less time consuming than having to watch and assess two videos per participant. Finally, it might have been less awkward for participants to record themselves when they knew that no one else was going to see their recordings. Normally, during longer training programmes, users get used to recording themselves after a while, but the experiment was too short for that. Therefore, self-assessment was chosen. On the other hand, having one coach assess all recordings might have led to more consistent results as some participants may be stricter than others. Participants, however, were likely consistent in the way they assessed themselves in both measurements. Presumably, this did not affect the learning outcome as this is person specific.

Both the baseline measurement and final measurement resulted into a score. Because everyone has communication skills to a greater or lesser extent, scores in TrainTool are around the number 100. Therefore, scores are always between 75 and 125 points. This can also be translated to a scale from 0 - 100. Participants received either a 100 or 0 points, depending on whether they met a criteria or not. The scores could be calculated by multiplying the correct amount of criteria by 100 and dividing this by the amount of criteria (6). These scores were translated to the TrainTool range to make them sound better for their users (e.g., receiving 33 points sounds worse than receiving 92). Participants could receive the following scores during the baseline and final measurement:

- 75 (0)
- 83 (17)
- 92 (33)
- 100 (50)
- 108 (67)
- 117 (83)
- 125 (100)

By subtracting the baseline score from the final score, the learning outcome could be measured (e.g., 125 (final) - 75 (baseline) = 50 (learning outcome)).

4.4.7 Survey

On the final page of the training, participants were thanked for their participation and asked to fill out a survey. This survey was made in Qualtrics, a survey tool used by Utrecht University. Because the survey data needed to be linked to the TrainTool data, participants first had to enter the email address they used to sign up for the training.

Next, they were asked about their demographics (e.g., their age, gender, educational attainment and employment status). Additionally, they were asked whether or not they had used TrainTool before.

The majority of the survey questions focused on how participants experienced the training. First, participants were presented questions about the quality of the training and the characters. Moreover, participants were asked to rate how easy it was to apply the theory and to which degree they learned something from the training.

Furthermore, 8 items from the User Experience Questionnaire (UEQ) were included in the survey to help evaluate the training. The UEQ is a fast and reliable questionnaire to measure the User Experience of interactive products (Schrepp et al., 2018). The full version of the UEQ contains 26 items, while the official shortened version contains only 8. An adjusted shortened version was used for this survey, selecting the items that seemed most relevant for evaluating the training.

Additionally, Tcha-Tokey et al. (2016) proposed and validated a unified questionnaire on User Experience in Immersive Virtual Environments. It contains 10 scales measuring, among other things, presence, engagement, immersion, flow, and usability. These scales were based on existing questionnaires and contained 87 items in total, most of which used a 10-point Likert scale. Several of these items were adapted to fit within the context of the training and its video characters.

Lastly, the survey was used to evaluate some of the characters that appeared in the training. In the original training programme, trainees are exposed to a variety of actors, differentiating in attributes, such as their gender, voice, ethnicity, and other physical attributes. Therefore, the deepfakes and virtual humans were attempted to resemble the original actors as closely as possible. Research has shown that how virtual humans are perceived and how effectively they facilitate learning can be influenced by their design (i.e., their voice, speech patterns, and physical appearance) (Ozogul et al., 2013; Schroeder et al., 2017). Therefore, four characters from the training were evaluated by including an image of them in the survey and asking participants to rate several statements about them (e.g., about their realism, appearance, credibility, voice, ...), using a 5-point Likert scale. This way, the perception of the same character could be compared between versions, providing insight into which voices were experienced as more natural or pleasant, and whether participants preferred certain physical attributes over others. Consequently, three versions of the same survey were created, each containing version appropriate images.

The entire survey can be found in Appendix B.

4.5 Procedure of the experiment

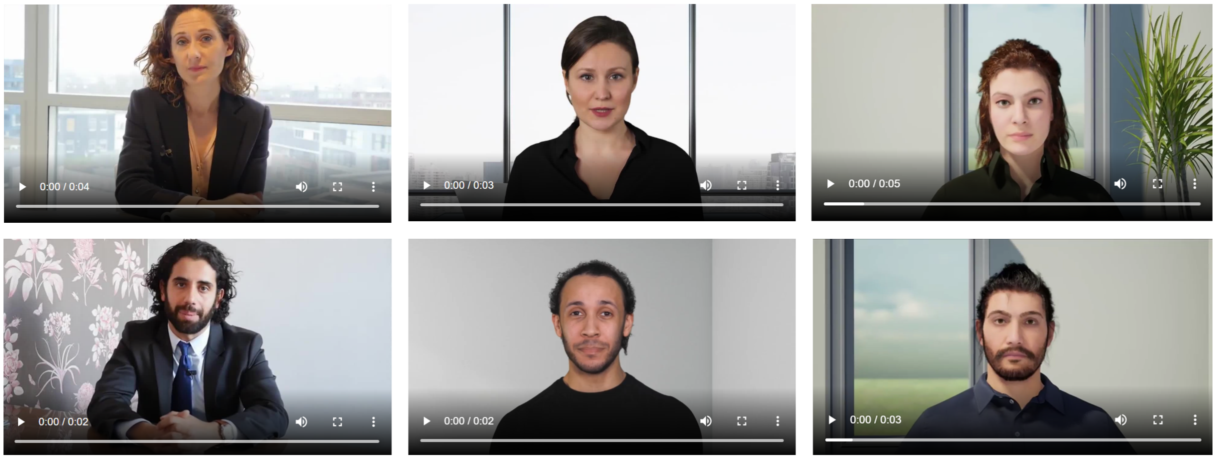


Figure 6: A comparison of two video characters between the different versions. From left to right: A) actors, B) deepfakes, C) virtual humans.

A between-subjects study design was used for the experiment. Participants were instructed to complete a short training programme and to fill out a survey afterwards. They were randomly assigned to one of the following groups: the A) actor version, B) deepfake version, or C) virtual human version, by sending them a version specific sign-up link. Via this link they read and signed the informed consent form and filled out their email address on the sign-up page. After this, they received a confirmation email allowing them to create a TrainTool account and go to the feedback training. Figure 6 shows two examples of what each version looks like.

After logging into TrainTool and reading the introduction, participants were instructed to do a practice role-play to set up their webcam and microphone. Before starting the actual training, a baseline measurement was done to find out how well they were able to respond to a situation without having practiced the skill yet. Next, participants had to practice this skill though three practice rounds. For each round, participants had to prepare by watching a theory video and one or two example videos (with corresponding questions). After the preparation, they were presented with a situation to which they had to respond. As soon as the video character finished talking, participants had to apply the theory and record themselves responding to the situation. Finally, after finishing the practice rounds, a final measurement was done. Participants were instructed to apply the skills they had learned and to respond to a situation that was similar to the baseline measurement. Finally, after submitting the recording, they were directed to the completion page of the training. On this page, which contained a link to Qualtrics, they were instructed to fill out the survey,

The entire experiment took participants between 20 and 30 minutes to complete. A schematic overview of the experiment is portrayed in figure 7.

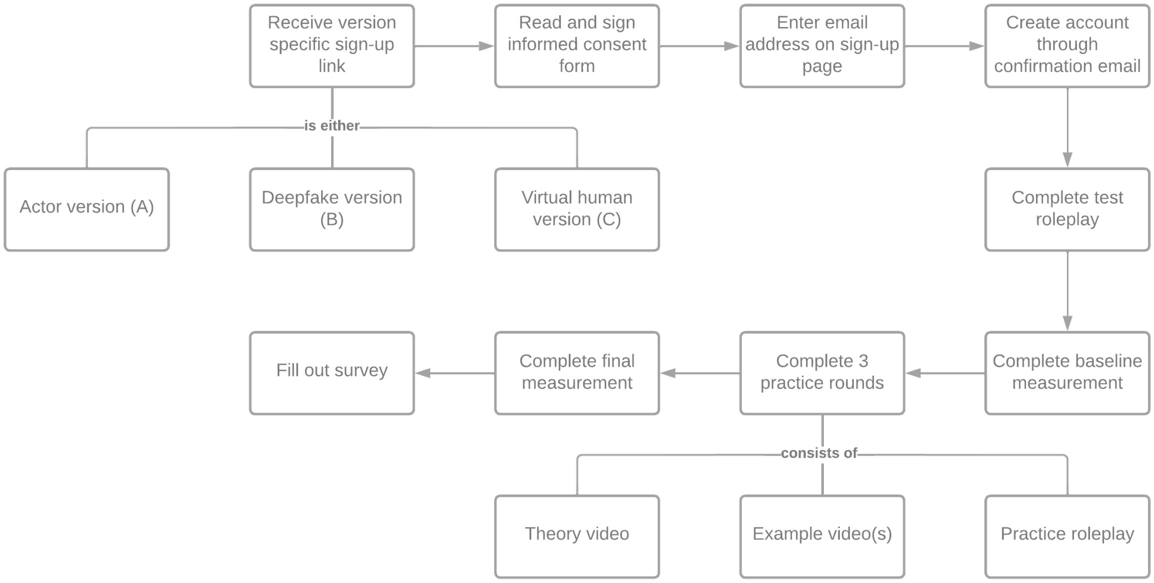

Receive version specific sign-up link
Read and sign informed consent form
Enter email address on sign-up page
Create account through confirmation email
is either
Actor version (A)
Deepfake version (B)
Virtual human version (C)
Complete test roleplay
Fill out survey
Complete final measurement
Complete 3 practice rounds
Complete baseline measurement
consists of
Theory video
Example video(s)
Practice roleplay

Figure 7: Schematic overview of the experiment.

4.6 Variables

The main independent variable in this study is the *character type* (actor, deepfake, virtual human). Other independent variables are *character attributes* (such as their gender, voice, or physical appearance), *participant demographics* (such as their age and gender), and the *time spent* to complete the training. These variables could have potentially influenced the dependent variables. For instance, they could help investigate whether there are any correlations between the age or gender of participants and how the characters are perceived by them (e.g., credibility, realism, discomfort, ...).

The first main dependent variable is *learning outcome*. This variable can be calculated by subtracting the *intake score* (baseline measurement) from the *outtake score* (final measurement). This can be compared to the *subjective learning outcome*, which participants were asked about in the survey (i.e., how much they think they have learned). To answer the research question, user experience had to be measured as well. For this, the scores from the survey's *UEQ* statements (Schrepp et al., 2018) were used. Additionally, to give more insight into the experience of users, variables such as *immersion* and *engagement* were analysed. The final dependent variable was *character perception*, consisting of, among other things, the perceived realism and credibility. These variables could provide insight into whether the perception of characters could be influenced by, for example, the character type or specific character attributes.

An overview of the main variables is presented in Table 1.

Independent variables	**Dependent variables**
Character type (actor, deepfake, virtual human)	Learning outcome (final score - intake score)
Character attributes (gender, voice, ...)	Subjective learning outcome
Participant demographics (age, gender, ...)	Immersion, engagement, presence
Time spent on training (in minutes)	Eight separate UEQ criteria
	Character perception (perceived realism, credibility, attractiveness, discomfort, ...)

Table 1: Overview of the variables.

The next chapter will elaborate upon the results from the experiment.

5 Results

This section will elaborate upon the data that was gathered and how it was analysed. Before the data was analysed, the TrainTool data (i.e., the intake scores and outtake scores) had to be combined with the Qualtrics survey data. There was a separate Qualtrics file for each version, so the three files were exported separately to SPSS, in which they were combined into one data file. A new 'version' variable was added to distinguish between the three versions. Finally, the TrainTool data, which was stored in an Excel file, was added to the SPSS file. Participants' email addresses were used as an identifier for correctly combining the data.

The first subsection will analyse the learning outcome (i.e., the outtake score - intake score). After that, the different components of the user experience will be elaborated upon and analysed. Lastly, the participants' perception of different video characters will be compared.

5.1 Learning outcome

A Shapiro Wilk normality test was done to determine whether the intake score, outtake score, and learning outcome were normally distributed or not. The test showed that the data significantly deviated from a normal distribution, as all p-values were below .05. Based on this outcome, a non-parametric test was used to further analyse the TrainTool data.

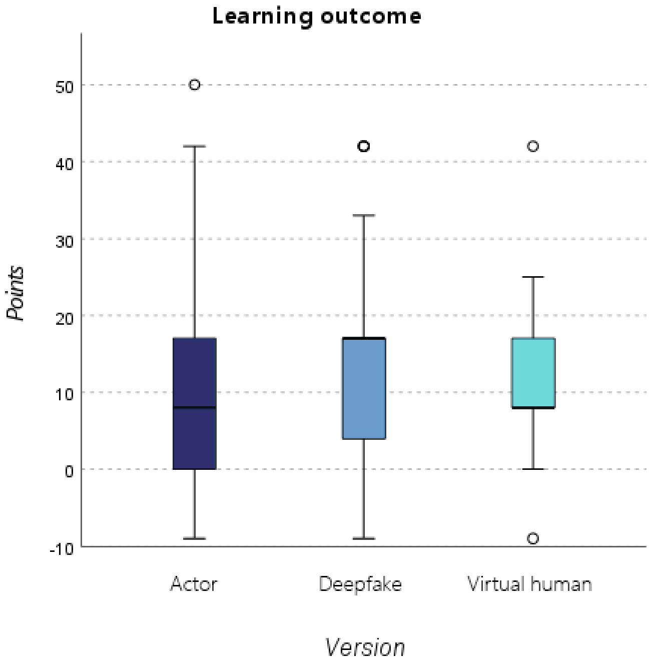


Figure 8: Boxplots comparing the learning outcome between versions.

The Kruskal Wallis H test showed that there was not a statistically significant difference in learning outcome between the different groups, $X^2(2) = 2.764$, $p = .251$, with a mean rank learning outcome of 46.67 for version A (actor), 58.15 for version B (deepfake), and 50.66 for version C (virtual human).

These results are visualised in the boxplots from Figure 8. The scale of the y-axis, the learning outcome in points, ranges from a negative outcome (-9 points) to the highest possible outcome (125 (max. score) - 75 (min. score) = 50 points). The negative outcome was obtained by a single individual that ended up with a slightly lower score than what they started with. For example, they made one mistake during the intake roleplay and two mistakes during the final roleplay. Possible explanations for this will be given in the discussion section.

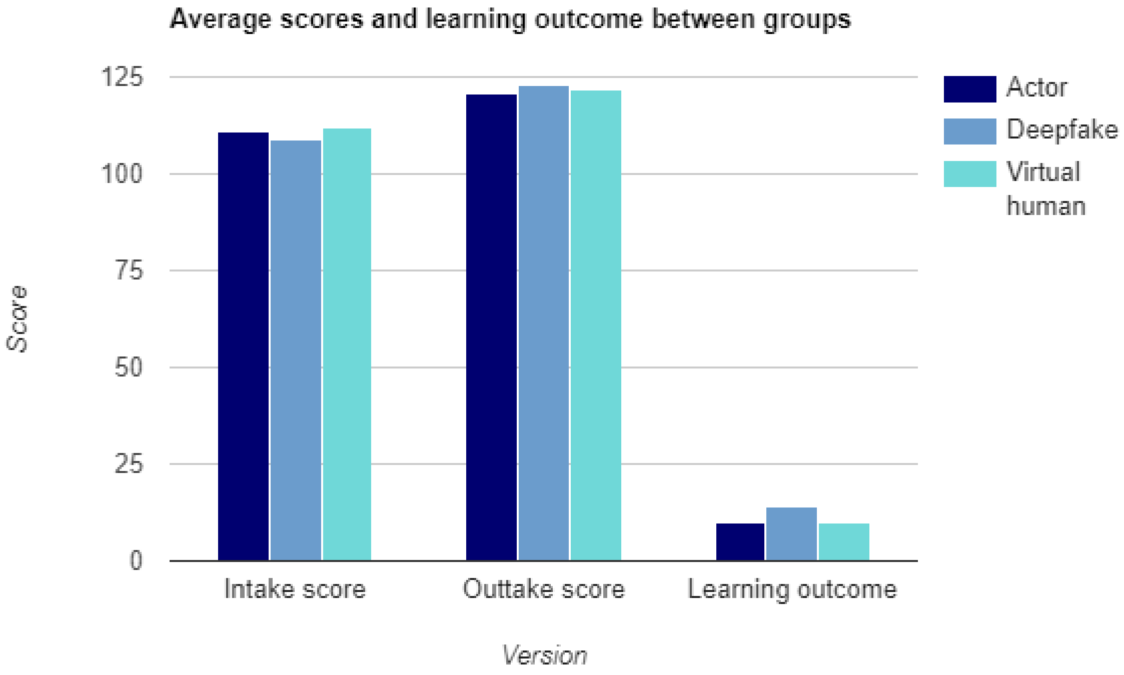


Figure 9: A bar chart comparing the mean scores and learning outcome between versions.

Additionally, by looking at the bar chart in Figure 9 it can be seen that the mean intake score (A: 111, B: 109, C: 112), outtake score (A: 121, B: 123, C: 122), and learning outcome (A: 10, B: 14, C: 10) do not differ much between the three versions. The Kruskal Wallis H test confirmed that these differences were not significant.

5.2 User experience

The user experience was measured through several survey questions, divided into separate categories. The survey contained questions about the quality of the training, some criteria from the User Experience Questionnaire (UEQ) (Schrepp et al., 2018), and some questions regarding immersion and presence in the training derived from research by Tcha-Tokey et al. (2016).

5.2.1 Quality of the training

Participants were asked to rate the following four statements about the quality of the training, using a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5):

1. I thought the quality of the training was good.
2. I thought the quality of the characters was good.
3. It was easy to apply the theory.
4. I have learned a lot from the training.

The Kruskal Wallis H test was used for each statement, of which two returned significant p-values.

First, the test showed that there was a statistically significant difference in the perception of the *quality of the training* between the different groups, $X^2(2) = 6.787$, $p = .034$, with a mean rank agreement of 47.74 for version A (actor), 61.35 for version B (deepfake), and 46.24 for version C (virtual human).

Moreover, the test showed that there was a statistically significant difference in the perception of the *quality of the characters* between the different groups, $X^2(2) = 17.048$, $p < .001$, with a mean rank agreement of 60.08 for version A (actor), 59.75 for version B (deepfake), and 35.96 for version C (virtual human).

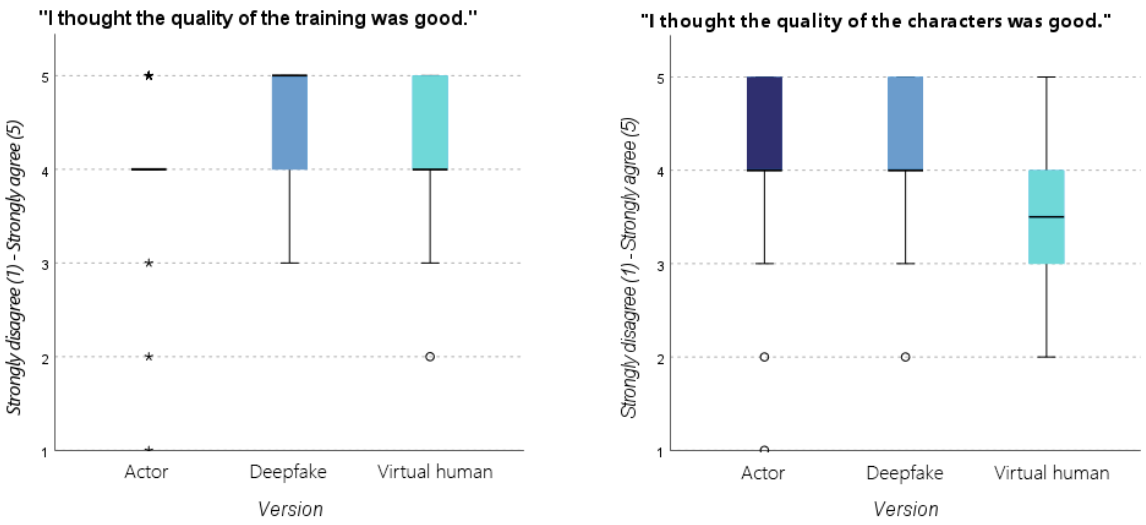


Figure 10: Boxplots comparing the quality of the training and characters between versions.

The boxplots in Figure 10 show how participants answered the statements about the quality of the training and the characters. The actor boxplot from the left image shows some extreme outliers and appears to have no 'box', however, the first (Q1) and third (Q3) quartile are both 4, just like the median. In addition to this flattened boxplot, the frequencies (in percentage) of each degree of agreement with the first statement are portrayed in Figure 11. This figure confirms that the majority of the actor group somewhat agreed (4) with the statement.

Dunn's test was used for the post hoc analysis to indicate between which groups these differences occurred. The test revealed that for the first statement significant differences were found between the virtual human and deepfake version (p = .018) and between the actor and deepfake version (p = .035). Thus, the quality of the training appeared to be rated significantly higher for the deepfake version than for the other two versions. However, when adjusted by the Bonferroni correction for multiple tests, these values were no longer significant (p = .055 and p = .105).

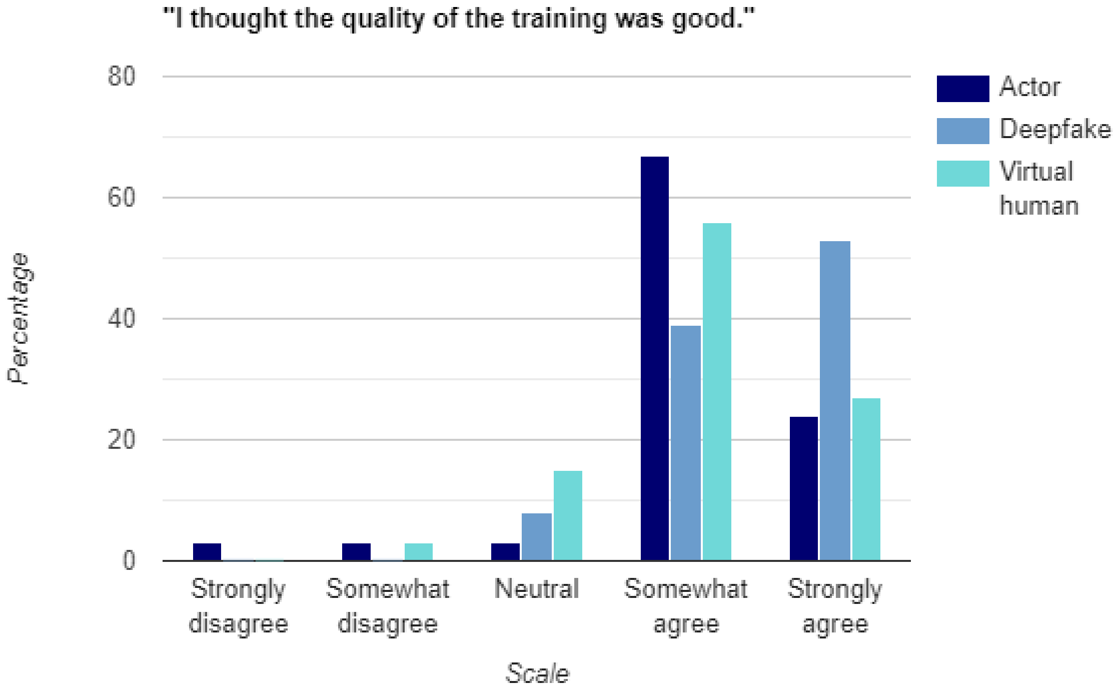


Figure 11: A histogram showing frequencies of the agreement (in percentage).

For the second statement, Dunn's test found significant differences between the virtual human version and deepfake version (p <.001) and between the virtual human version and actor version (p <.001). Adjusted by the Bonferroni correction these values remained significant, both with a significance of p = .001. Thus, the quality of the virtual human characters was perceived significantly lower than the quality of the other two versions.

For the latter two statements no significant differences between versions were found. This falls in line with the learning outcome results, in which no significant differences were found either. Presumably, how easy it was to apply the theory and how much participants indicated they had learned from the training could be related to the learning outcome.

5.2.2 UEQ criteria

Participants were asked to rate the training based on the following criteria, using a 7-point Likert scale:

1. Annoying - Enjoyable
2. Complicated - Easy
3. Unpleasant - Pleasant
4. Not interesting - Interesting
5. Conventional - Inventive
6. Unattractive - Attractive
7. Bad - Good
8. Usual - Leading Edge

The Kruskal Wallis H test showed significant values for two of the criteria, namely *conventional - inventive* and *usual - leading edge.*

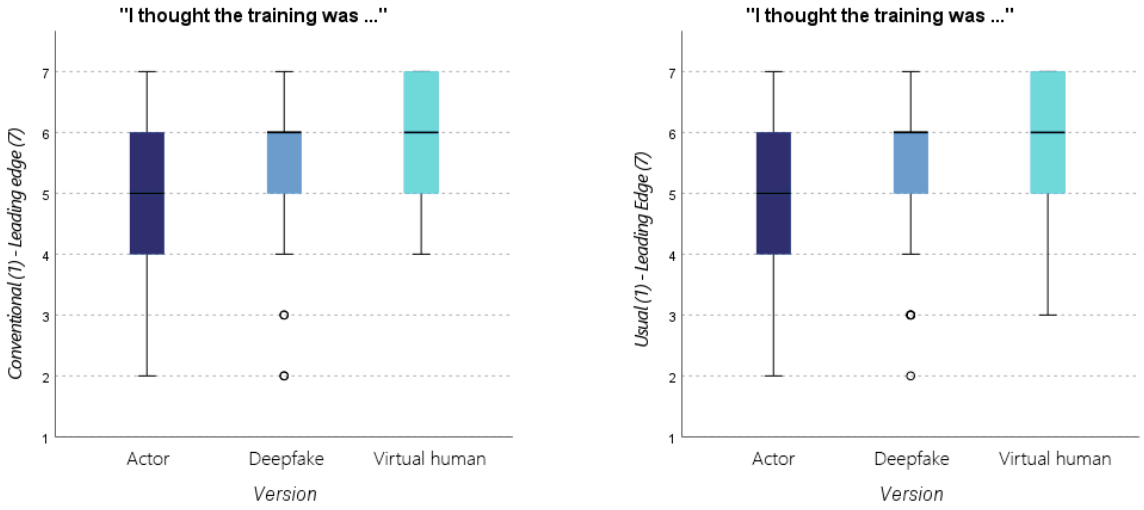


Figure 12: Boxplots of the UEQ criteria that differ significantly between versions.

First, the test showed that there was a statistically significant difference in how *conventional or inventive* the training was found by the different groups, $X^2(2) = 6.099$, p

= .047, with a mean rank of 42.29 for version A (actor), 53.97 for version B (deepfake), and 59.34 for version C (virtual human).

Additionally, the test showed that there was a statistically significant difference in how *usual or leading edge* the training was found by the different groups, $X^2(2) = 6.474$, p = .039, with a mean rank of 42.89 for version A (actor), 51.96 for version B (deepfake), and 60.88 for version C (virtual human).

Dunn's test revealed that for both criteria the differences were found between the actor version and the virtual human version, with a significance of p = .016 and p = .011. Adjusted by the Bonferroni correction, the significance became p = .047 for *conventional - inventive* and p = .033 for *usual - leading edge*. These differences are visualised in the boxplots from Figure 12.

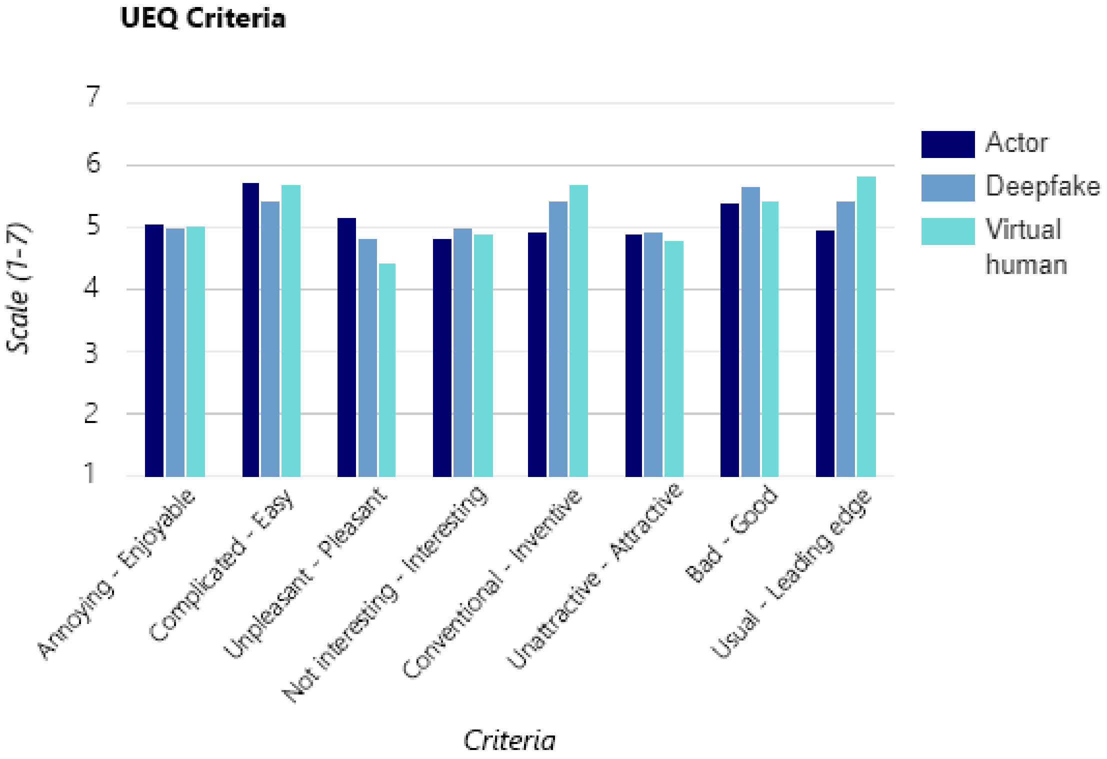


Figure 13: A bar chart comparing the mean ratings of the UEQ criteria.

There were no significant differences found for the remaining six criteria, suggesting that the character type did not have a large influence on how enjoyable, pleasant, interesting, attractive, or good the training was perceived by participants. The mean ratings of the UEQ criteria are portrayed in Figure 13.

5.2.3 Immersion, engagement and presence

Next, participants were presented with several statements related to immersion, engagement and presence, which they had to rate using a 10-point Likert scale ranging from strongly disagree (1) to strongly agree (10):

1. My interactions with the role-play partners seemed natural. *(presence)*
2. The visual aspects of the role-play partners involved me. *(engagement)*
3. I was involved in the role-play experience. *(engagement)*
4. I felt stimulated by the role-play/ training environment. *(immersion)*
5. I became so involved in the training environment that I was not aware of things happening around me. *(immersion)*

The Kruskal Wallis H test showed no statistically significant differences for any of the statements between the three versions. This is visualised in Figure 14, which only shows small differences in the degree of agreement with each of the five statements.

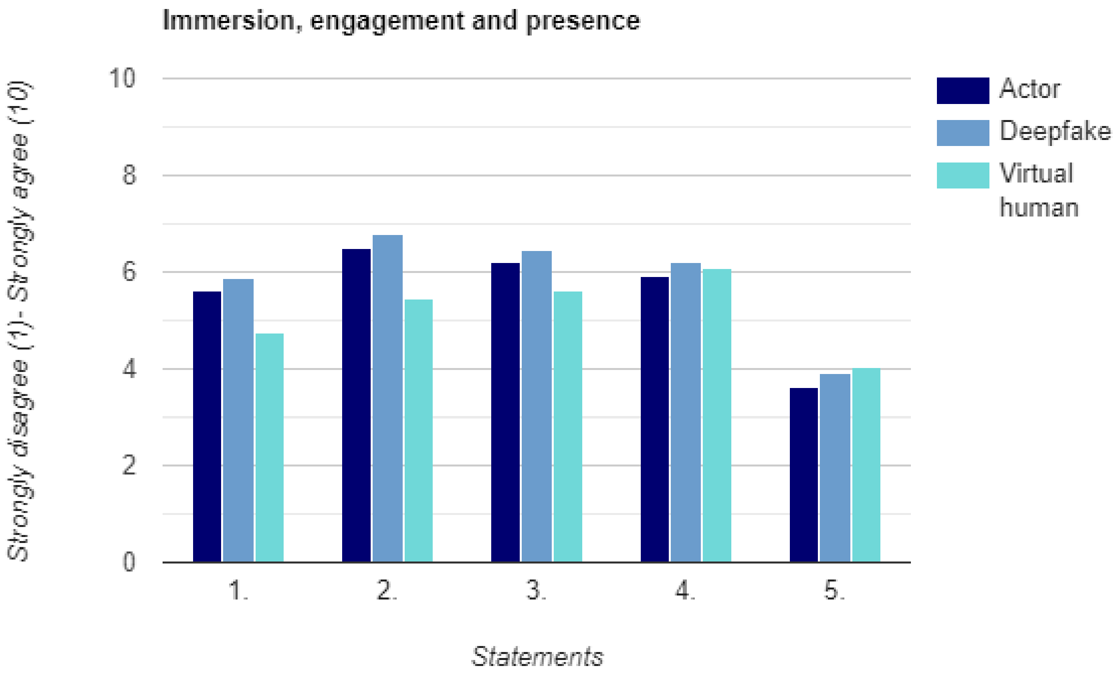


Figure 14: A bar chart comparing the mean ratings of the statements regarding immersion, engagement and presence.

5.3 Character perception

Finally, participants were instructed to rate four different characters (2 male, 2 female) based on the following statements, using a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5):

1. I thought he/she was realistic.

2. I felt discomfort looking at him/her.
3. I thought he/she was credible.
4. I thought he/she was sympathetic.
5. His/her physical appearance looks appealing.
6. I felt engaged watching and listening to him/her.
7. His/her voice sounded natural.
8. His/her voice sounded pleasant.

To analyse these statements the Kruskal Wallis H test was used, followed by Dunn's test for the post hoc analysis.

5.3.1 Male character 1

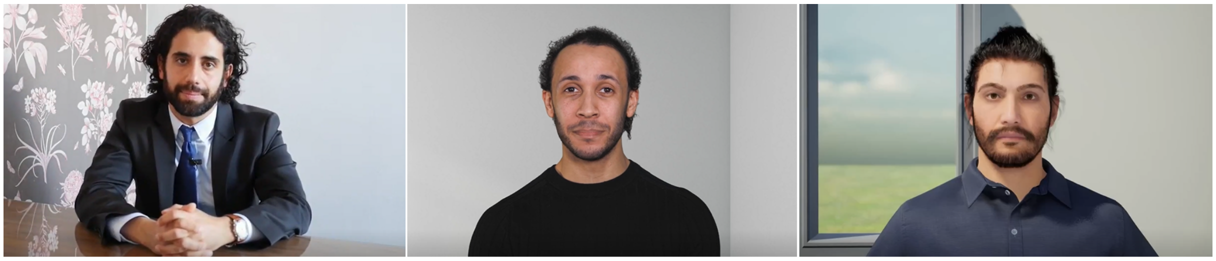


Figure 15: Male character from the intake and outtake roleplay.

For the first male character, portrayed in Figure 15, significant differences were found for statement 4 (p <.001) and 7 (p = .010).

For the fourth statement, the virtual human was rated less sympathetic than the actor (p = .019) and deepfake (p <.001). The Bonferroni correction adjusted these values into p = .058 and p = .000 respectively, implicating that the virtual human only differentiates significantly from the deepfake and not from the actor. These differences are visualised in Figure 16. The boxplot also seems to show that the deepfake was perceived more sympathetic than the actor, however, this difference was not significant (p = .105).

For the seventh statement, the virtual human's voice was rated less natural than the actor's voice (p = .018, adjusted: p = .053) and deepfake's voice (p = .005, adjusted: p = .014). This can also be seen in Figure 16. It makes sense that there was no significant difference between the actor's and deepfake's voice, since the same audio file was used for both versions.

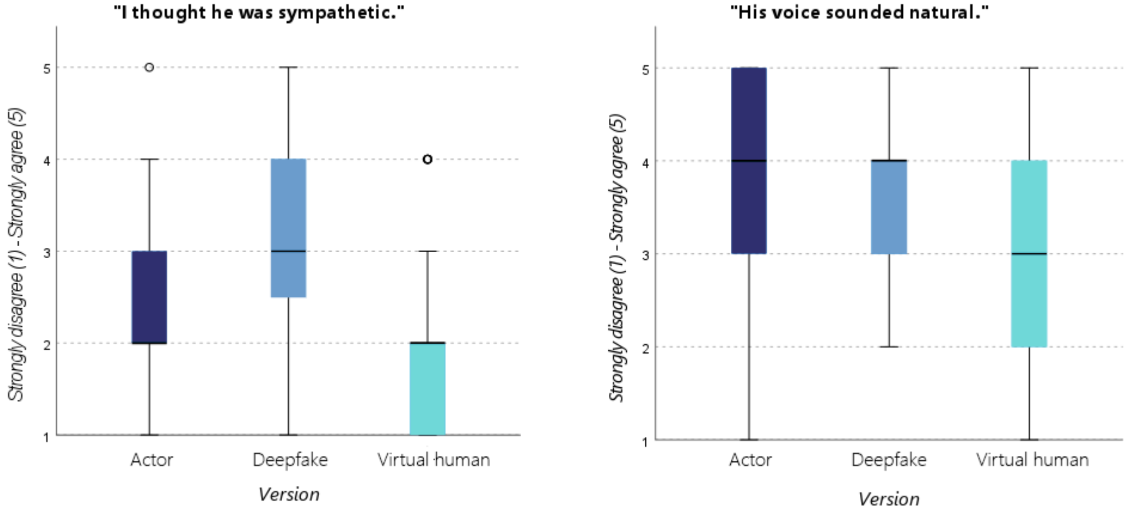


Figure 16: Boxplots of the differences in perception of the first male character.

5.3.2 Male character 2

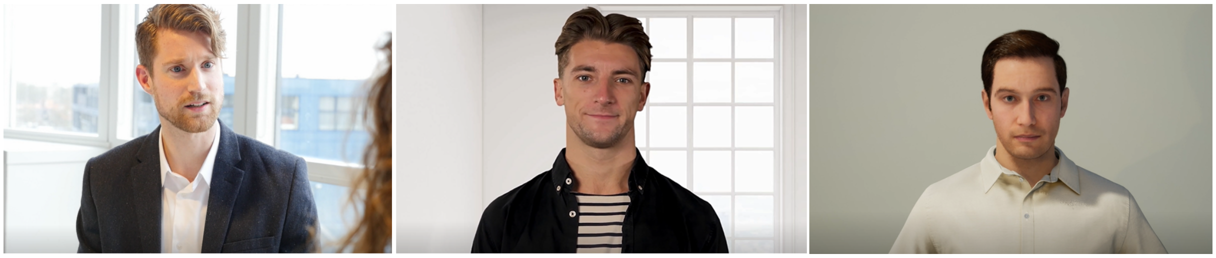


Figure 17: Male character from the example videos.

For the second male character, portrayed in Figure 17, significant differences were found for statement 1, 3, 4, 5, 7, and 8. Table 2 contains the significant values from the post hoc analysis. The bold values between the parentheses are adjusted by the Bonferroni correction. These adjusted values suggest that there was no significant difference in the perception of the character's voice. Most differences were found between the virtual human and deepfake character. All differences are visualised in the boxplots from Figure 18.

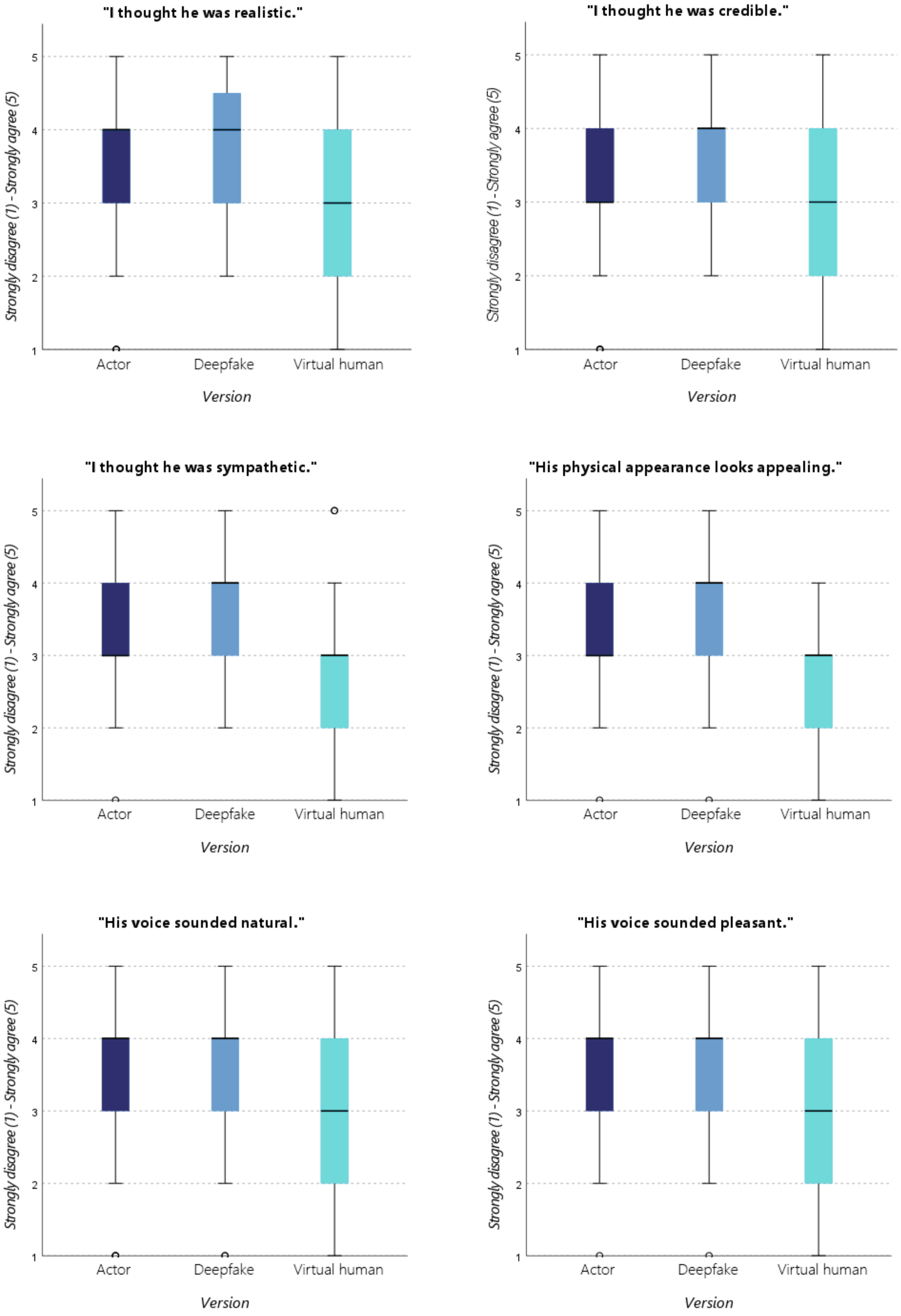


Figure 18: Boxplots of the differences in perception of the second male character.

	Virtual human - Actor	***Virtual Human - Deepfake***	***Actor - Deepfake***
1. Realistic		p = .001 **(p = .004)**	
2. Discomfort			
3. Credibility		p = .016 **(p = .047)**	p = .049 (p = .146)
4. Sympathetic	p = .022 (p = .066)	p = .002 **(p = .005)**	
5. Appealing	p <.001 **(p = .003)**	p <.001 **(p = .000)**	
6. Engagement			
7. Voice (natural)		p = .021 (p = .063)	
8. Voice (pleasant)	p = .025 (p = .075)	p = .021 (p = .062)	

Table 2: Significant differences in perception of the second male character. Adjustments by the Bonferroni correction between the parentheses.

5.3.3 Female character 1

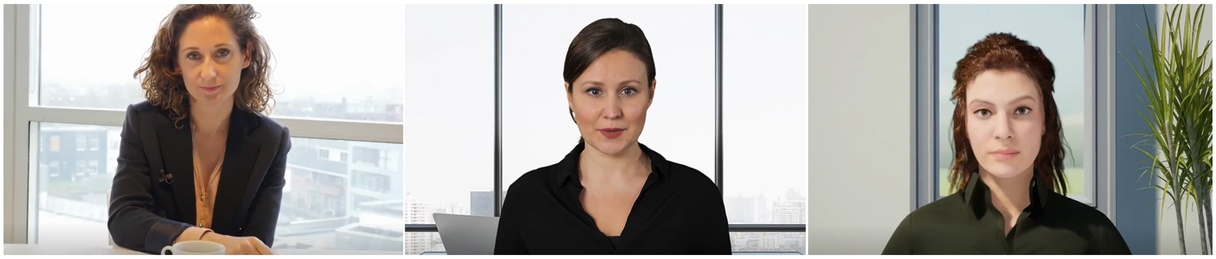


Figure 19: Female character from the practice roleplays.

For the first female character, portrayed in Figure 19, significant differences were found for statement 1 (p = .002) and 7 (p = .003). These differences are visualised in Figure 20.

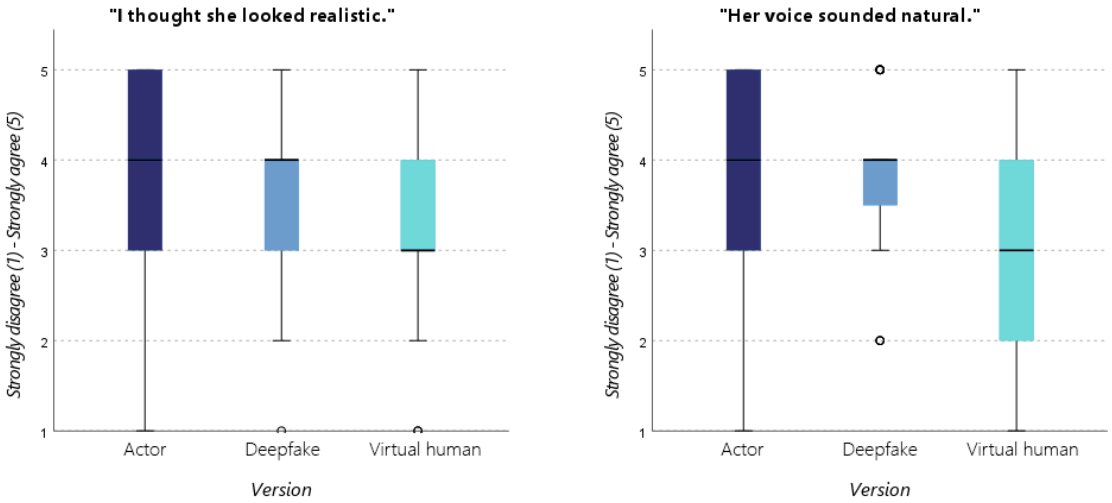


Figure 20: Boxplots of the differences in perception of the first female character.

Dunn's test showed that the virtual human was perceived less realistic than the actor (p <.001, adjusted: p = .001) and than the deepfake (p = .034, adjusted: p = .101). There was no significant difference between the actor and deepfake.

Additionally, the virtual human's voice was found to sound less natural than the actor's (p = .003, adjusted: p = .008) and deepfake's voice (p = .004, adjusted: p = .013).

5.3.4 Female character 2

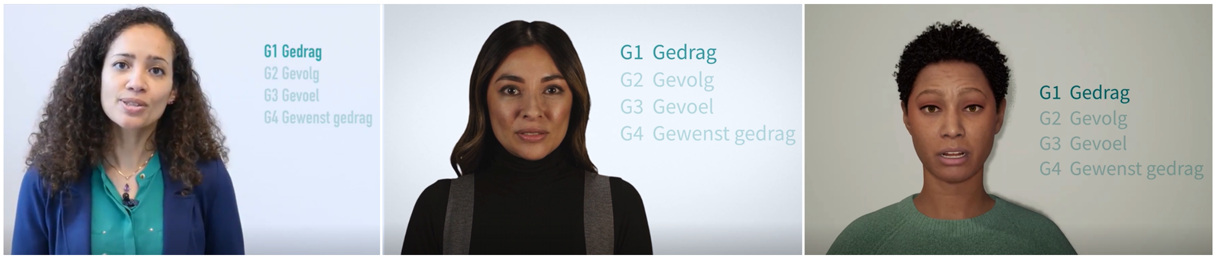


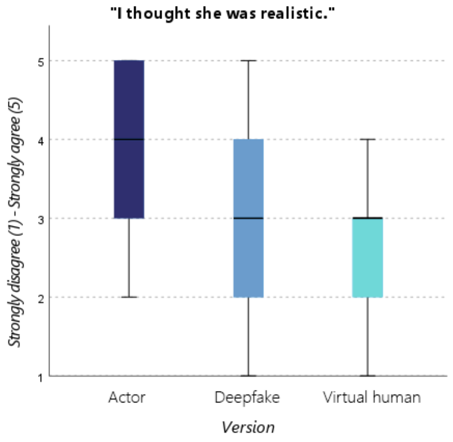
Figure 21: Female character from the theory videos.

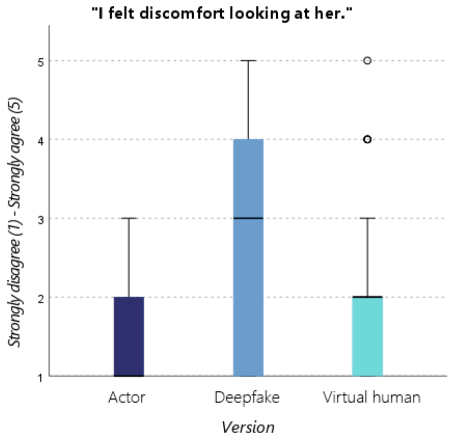
For the second female character, portrayed in Figure 21, significant differences were found for all statements. Table 3 contains the significant values from the post hoc analysis. Again, most differences were found between the actor and virtual human, however, Dunn's test also found significant differences between the actor and deepfake. These initial significance values were also adjusted by the Bonferroni correction, found between the parentheses. The bold values remained significant after the correction. All differences are visualised in Figure 22.

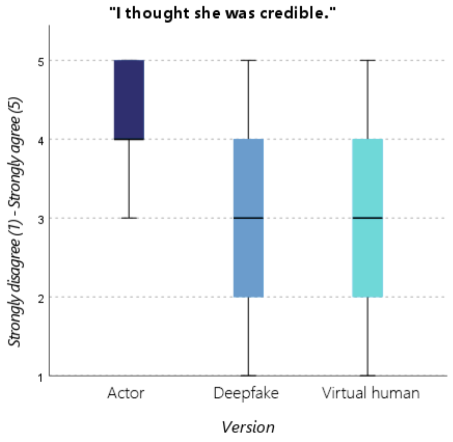
	Virtual human - Actor	***Virual Human - Deepfake***	***Actor - Deepfake***
1. Realistic	p <.001 **(p = .000)**		p <.001 **(p = .000)**
2. Discomfort		p = .035 (p = .106)	p <.001 **(p = .000)**
3. Credibility	p <.001 **(p = .000)**		p <.001 **(p = .000)**
4. Sympathetic	p <.001 **(p = .000)**		p <.001 **(p = .000)**
5. Appealing	p = .001 **(p = .004)**	p = .003 **(p = .010)**	
6. Engagement	p = .008 **(p = .025)**		p = .022 (p = .065)
7. Voice (natural)	p <.001 **(p = .000)**	p = .002 **(p = .007)**	
8. Voice (pleasant)	p <.001 **(p = .000)**	p = .027 (p = .080)	

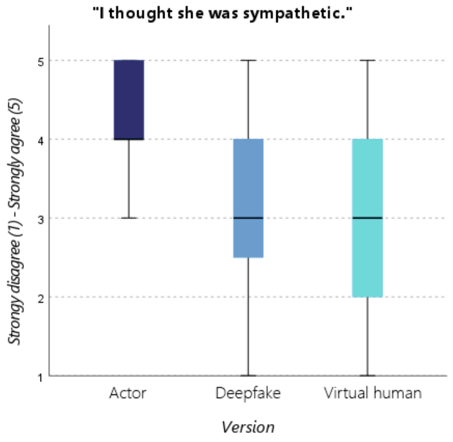
Table 3: Significant differences in perception of the second female character. Adjustments by the Bonferroni correction between the parentheses.

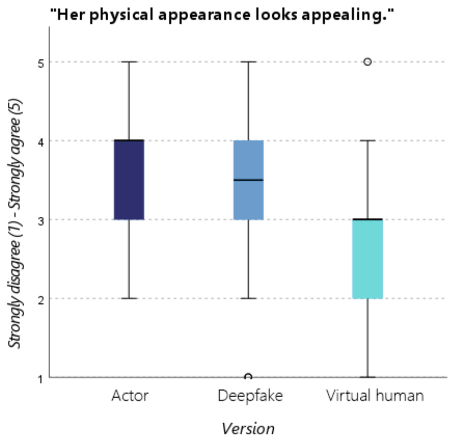
This character explained the theory to participants through three separate videos, each lasting around 30 seconds to one minute. These videos were quite long compared to the actual roleplay videos (i.e., male character 1 and female character 1), which typically last between 3 to 10 seconds. Therefore, participants were watching the theory character a lot longer, giving them more time to observe her. This could explain why the deepfake and virtual human were often perceived worse than the actor. Additionally, the avatar choice for the deepfake and the design of the virtual human might have influence on the way it is perceived. The next section will elaborate on this further.

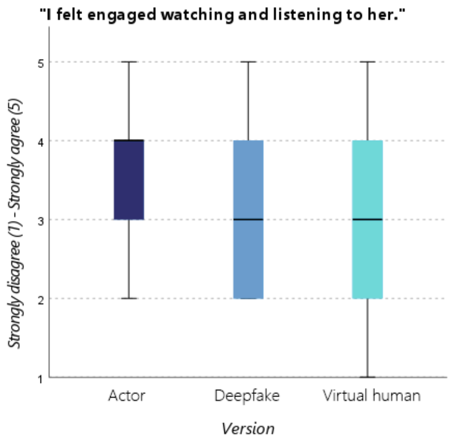

"I thought she was realistic."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version


"I felt discomfort looking at her."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version


"I thought she was credible."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version


"I thought she was sympathetic."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version


"Her physical appearance looks appealing."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version


"I felt engaged watching and listening to her."
Strongly disagree (1) - Strongly agree (5)
1
2
3
4
5
Actor
Deepfake
Virtual human
Version

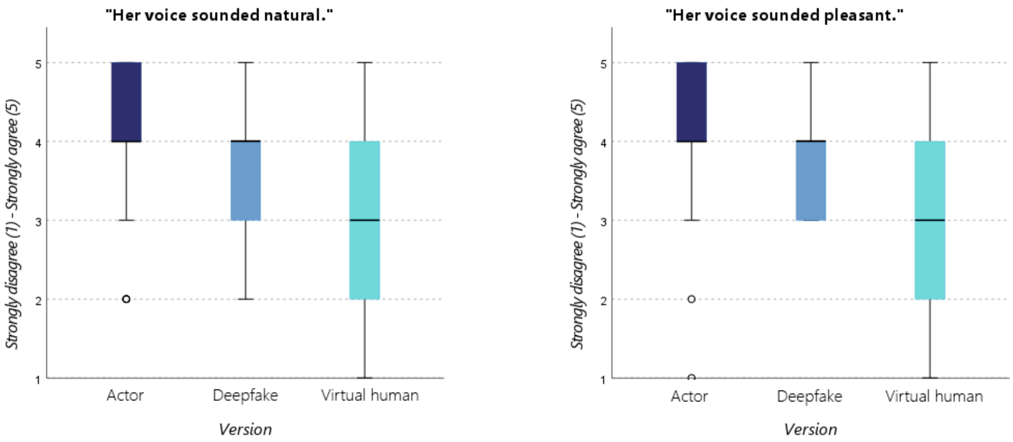


Figure 22: Boxplots of the differences in perception of the second female character.

6 Discussion

This chapter will discuss the key findings from the results chapter in relation to the research aims and research questions. First, a quick overview of the study will be given, after which the results will be interpreted. Moreover, the limitations of the study will be reviewed and opportunities for future research will be proposed.

6.1 Overview of the study

This research was conducted for Faculty of Skills, a training agency that offers a variety of soft skill training programmes to improve place communication. To train these skills, trainees participate in several video role-plays in which they act out real-life situations. They watch a pre-recorded video, which is created in a studio with actors, and reply to it once the video is finished. Because creating these videos is a time-consuming and expensive process, two alternative, digital actors were explored, namely deepfakes and virtual humans.

In order to answer the main research question, *"How do different character representations in a virtual communication training influence the experience and learning outcome of trainees?"*, a between-subjects study was designed. Participants were randomly assigned to either version A) actors, B) deepfakes, or C) virtual humans. It was hypothesized that there would be no significant difference in learning outcome and experience between the versions. For the experiment, participants completed a (shortened) training programme, which was used to measure their learning outcome. Afterwards, they filled out a survey, which mainly focused on their experience during the training.

Previous studies have extensively researched virtual humans as agents in virtual (training) environments, however, not in the context of pre-recorded videos, as participants often directly interacted with them in the environment. Additionally, not much research had been found on deepfakes in the context of (communication) training, as most research focused on the drawbacks and dangers of deepfakes, rather than practical, positive applications. Therefore, the results from the experiment might help bridge this gap in literature.

The results were analysed in the previous chapter and will be further interpreted in the following section.

6.2 Interpreting the results

6.2.1 Learning outcome

No significant differences were found in learning outcome between the three groups, suggesting that the type of character representation in the video did not influence how much participants had learned from the training.

As stated in the results chapter and shown in Figure 8, the lowest learning outcome that was obtained was a negative score, namely -9. In total, there were four participants that obtained a negative learning outcome. A negative outcome could have occurred when they, for example, initially only failed to meet one criteria during the intake, but ended up missing two criteria during the outtake. An explanation for this could be that they might have been too focused on including everything they had learned, making their answer less concise, which was one of the assessment criteria. Another explanation could be that some people assess themselves more rigorously than others. These cases were not removed from the data set, because every group contained at least one person with a negative score. Moreover, their experience during the training was still a valuable addition to the data.

6.2.2 User experience

The user experience was measured through a variety of questions. Some focused on how participants experienced the quality of the training and characters, while the UEQ criteria focused more on other components of the user experience, such as how enjoyable, pleasant, or interesting they found the training.

The quality of the training appeared to be rated significantly higher for the deepfake version, however, after applying the Bonferroni correction of multiple tests, these values were no longer significant. This suggests that the type of character representation did not influence participants' perception of the quality of the training. The quality of the characters, on the other hand, was experienced differently among the groups. More specifically, the quality of the virtual humans was rated significantly lower than the quality of the other two character types.

Nevertheless, the virtual human training was rated as more inventive and leading edge than the other two versions. Other than that, no significant differences were found in the remaining UEQ criteria, suggesting that participants did not have a significantly different experience during the training, at least not based on e.g., how enjoyable, interesting, or attractive they found the training.

6.2.3 Character perception

Several significant differences were found in the perception of characters between groups. In the methodology chapter, two sub research questions were formulated regarding the influence of the character's gender and voice. The results showed significant differences in how natural and sometimes pleasant the voices were experienced. More specifically, the virtual humans were always rated worse on these criteria than the actors and deepfakes. There were no significant differences between the voices of the actors and deepfakes, which was expected since they both used the same audio file. The virtual humans, on the other hand, spoke in an AI voice that was created via a TTS tool. Unfortunately, the Dutch TTS voices still sound less natural than the English voices, which will be further discussed in the limitations of the study.

It seems like the character's gender does not influence the way it is perceived, as there were one male and one female character that were rated quite positively, and one male and one female character that were rated quite negatively. Again, most of the differences occurred between the virtual human and actor, and sometimes the virtual human and deepfake, with the virtual human always being rated worse than the other types.

The characters with the most differences between groups were the ones that appeared in the longest videos (i.e., the theory and example videos). Therefore, video length might influence the way characters are perceived. The longer participants are exposed to a video character, the more time they have to critically watch them. Additionally, participants did not have to think about what to say next, like they do when watching the actual (shorter) role-play videos. This gave participants more time to notice things that were 'off' about them, such as delayed lip movements, which could explain why these characters were rated worse than the other two.

Moreover, the style choices made when designing or choosing the character could have influenced the results. For instance, the second female character (from the theory videos) was rated rather poorly. Not just the virtual human, as the deepfake was also rated significantly worse than the actor. In addition to the video length, the deepfake character had quite a penetrating, intense look, which could explain why participants felt uncomfortable watching her and found her appearance less appealing. Lastly, the character's attitude (i.e., the role they played) could have influenced the way they were perceived. For instance, the character from the intake and outtake was generally not rated very sympathetic because he reacted quite bluntly and curtly. The deepfake character was rated slightly higher, although not significantly, but this could have been because he looked friendlier than the other two. Therefore, the avatar choice and character design, in addition to the video length, seem to have a larger influence on the way the characters are perceived than for example their gender does.

6.3 Limitations of the study

6.3.1 Quality of the virtual humans

The quality of the virtual humans was not very high. This was mainly due to the available soft- and hardware, as well as a lack of knowledge of and experience with the MetaHumans and Unreal Engine. First of all, Unreal Engine requires a high spec machine, specifically when working with large files that contain MetaHumans. Luckily a slightly better laptop could be borrowed, but the film-like quality as can be seen in YouTube videos could unfortunately not be achieved. To improve the lip sync and to add body motion, which would make the MetaHuman look less stiff, there is more paid software and hardware available, such as iClone or mocap. For the experiment, the bare minimum was used, which is an iPhone with a TrueDepth camera. Unfortunately, this was not already in possession and acquaintances did not have a spare one to borrow either. So, an iPhone 11 was bought in order to create the MetaHuman videos. It was still quite hard to correctly sync the lip movements while listening to the TTS audio file. This was slightly improved during editing by speeding up or delaying the video to match better with the audio. However, it was still not perfect, which could have been especially noticeable in the longer videos.

Furthermore, the lack of knowledge and experience made it very time-consuming to make the MetaHuman videos. A lot of documentation was read and tutorials were watched, but this self-thought knowledge did not result in the high quality videos from these tutorials. Because it was so time-consuming, at some point just settling for less was necessary in order to continue with the experiment.

Because creating these videos was time-consuming and often frustrating due to e.g., lags, it will definitely not be an easier or quicker solution for Faculty of Skills at this point in time. Additionally, the quality of the Dutch TTS voices is not as good as the English voices that are available. There are often fewer voices available and they also sound less natural. Even after spending a lot of time tweaking them, they were still not up to par. Therefore, using actual human voices is currently preferred over using TTS voices, as they sound more natural and pleasant. This especially goes for deepfakes, as the contrast is even bigger between a real-looking person and a fake-sounding voice.

6.3.2 Sampling and distribution

Although the sample size was already quite large, a better distribution of gender would have been preferred. Participants were randomly selected to a version, but the current distribution is very skewed, which made it impossible to analyse the data based on the participants' gender. This could have been prevented by using an additional sampling method, such as, purposive or quota sampling to make sure each group had about the same number of men and women in it. Additionally, it was difficult to find enough people over the age of 30 to participate, as fellow students were the easiest to recruit.

6.3.3 Training length and assessment

The training from the experiment was a lot shorter than an actual training would be. It might have been more representative to conduct this experiment with a full-length training, however, that would have also made it very difficult to find enough participants.

Furthermore, self-assessment was chosen for the experiment. There was no way to check if people seriously participated and assessed themselves. Although, according to Faculty of Skills, experience showed that people are perfectly capable of assessing themselves, having one coach assess all videos might result into more consistent assessments.

6.4 Future research

Building on the previous section, it might be insightful to investigate whether the results still hold up during a regular, longer training. Additionally, a similar experiment could be conducted using a training containing more emotions, such as how to deal with aggression. It would be interesting to find out whether these emotions can be translated properly via a deepfake or virtual human.

Furthermore, one might investigate how the gender or age of participants influences the way they perceive the training and characters. Unfortunately, the majority of people that participated in this research were male students, so the age and gender distribution was quite skewed. Therefore, due to the current data, this could not be analysed.

Finally, Faculty of Skills could conduct a pilot study in which they offer a tailor-made training to one of their customers. The Synthesia deepfakes seem, based on the findings from this study, to be a promising alternative for human actors. They may not be suitable for longer videos like the theory or example videos, but the actual role-play characters did not differ significantly from their actor counterparts. Faculty of skills offers tailor-made training programmes to their customers. The structure and theory of such a training is the same as in the original one, however, the role-plays can be personalised to specifically fit the client. Synthesia would allow Faculty of Skills to quickly create tailor-made programmes for their clients, which they could offer at a lower price since this version is cheaper and quicker to make. Instead of casting and hiring actors and spending the day in a studio, they would only have to reach out to a voice actor. Creating these audio files will be quicker and cheaper, and once obtained they can be easily uploaded to Synthesia with an avatar of choice. Adding these Synthesia role-play videos to the already existing theory and example videos would make the creation of tailor-made programmes a lot easier, quicker, and less expensive, which makes this a promising opportunity to investigate further.

7 Conclusion

Finally, the last chapter will conclude this thesis by answering the main research question.

To answer the main research question, *"How do different character representations in a virtual communication training influence the experience and learning outcome of trainees?"*, the following null hypotheses were formulated:

- H0: There is no significant difference in learning outcome between different character types.
- H0: There is no significant difference in experience between different character types.

Based on the results and discussion, the first null-hypotheses will be accepted, as the Kruskal Wallis H test found no significant difference in learning outcome between the three groups.

The second null-hypotheses will be rejected, because the survey results suggest that there are in fact some differences in the experience. More specifically, these differences are found between the actor and virtual human version and between the deepfake and virtual human version. No differences were found between the actor and deepfake version, except for some specific character perceptions.

This study aimed to investigate whether human actors in role-play videos could be replaced by digital alternatives. The results suggest that deepfakes could replace actors in short role-play videos, but possibly not in longer videos, such as theory videos. Virtual humans, on the other hand, do currently not seem like a viable option, specifically because the quality is perceived worse than the quality of the deepfakes and actors.

In conclusion, learning outcome does not seem to be influenced by the type of character representation. However, some components of the user experience are. More specifically, virtual humans, although rated more inventive, are also of lesser quality and simply not up to par compared to the quality of the deepfakes and human actors. Because no significant differences between deepfakes and actors were found, deepfakes could become a promising opportunity for Faculty of Skills to explore further. Keeping video length and avatar choice in mind, a first pilot application could be via a tailor-made training, adding the Synthesia videos to already existing theory and examples.

References

Aspegren, K. (1999). Beme guide no. 2: Teaching and learning communication skills in medicine-a review with quality grading of articles. *Medical teacher*, *21*(6), 563–570.

Baylor, A. L. (2005). Preliminary design guidelines for pedagogical agent interface image. In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 249–250).

Boneh, D., Grotto, A. J., McDaniel, P., & Papernot, N. (2020). Preparing for the age of deepfakes and disinformation. *HAI Policy Brief*.

Bosse, T., Gerritsen, C., & de Man, J. (2016). An intelligent system for aggression de-escalation training. In *Ecai 2016* (pp. 1805–1811). IOS Press.

Castillo, S., Hahn, P., Legde, K., & Cunningham, D. W. (2018). Personality analysis of embodied conversational agents. In *Proceedings of the 18th international conference on intelligent virtual agents* (pp. 227–232).

Chawla, R. (2019). Deepfakes: How a pervert shook the world. *International Journal of Advance Research and Development*, *4*(6), 4–8.

Chetty, G., & White, M. (2019). Embodied conversational agents and interactive virtual humans for training simulators. In *Proc. the 15 th international conference on auditory-visual speech processing* (pp. 73–77).

Cowell, A. J., & Stanney, K. M. (2005). Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International journal of human-computer studies*, *62*(2), 281–306.

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, *114*, 193–205.

Craig, S. D., & Schroeder, N. L. (2018). Design principles for virtual humans in educational technology environments. In *Deep comprehension* (pp. 128–139). Routledge.

Craig, S. D., Twyford, J., Irigoyen, N., & Zipp, S. A. (2015). A test of spatial contiguity for virtual human's gestures in multimedia learning environments. *Journal of Educational Computing Research*, *53*(1), 3–14.

Day, C. (2019). The future of misinformation. *Computing in Science & Engineering*, *21*(01), 108–108.

Draude, C. (2011). Intermediaries: reflections on virtual humans, gender, and the uncanny valley. *AI & society*, *26*(4), 319–327.

Faculty of Skills. (2021a). *About faculty of skills.* Retrieved from `https://www.faculty.nl/en/about`

Faculty of Skills. (2021b). *Trainingen.* Retrieved from `https://www.faculty.nl/nl/trainingen`

Faculty of Skills. (2021c). *Traintool.* Retrieved from `https://www.faculty.nl/nl/traintool-app/`

Faculty of Skills. (2022). *Data security faq - traintool.* Retrieved from `https://www.faculty.nl/en/about`

Fang, Z., Cai, L., & Wang, G. (2021). Metahuman creator the starting point of the metaverse. In *2021 international symposium on computer technology and information science (isctis)* (pp. 154–157).

Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, *70*(4), 455–471.

Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence. *Human–Computer Interaction*, *30*(5), 401–432.

Fox, J., Bailenson, J. N., & Tricase, L. (2013). The embodiment of sexualized virtual selves: The proteus effect and experiences of self-objectification via avatars. *Computers in Human Behavior*, *29*(3), 930–938.

Gartmeier, M., Bauer, J., Fischer, M. R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., ... Prenzel, M. (2015). Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. *Instructional Science*, *43*(4), 443–462.

Haake, M., & Gulz, A. (2009). A look at the roles of look & roles in embodied pedagogical agents–a user preference perspective. *International Journal of Artificial Intelligence in Education*, *19*(1), 39–71.

Hays, M. J., Campbell, J. C., Trimmer, M. A., Poore, J. C., Webb, A. K., & King, T. K. (2012). *Can role-play with virtual humans teach interpersonal skills?* (Tech. Rep.). UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES INST FOR CREATIVE TECHNOLOGIES.

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with computers*, *18*(2), 227–245.

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021). Countering malicious deepfakes: Survey, battleground, and horizon. *arXiv preprint arXiv:2103.00218*.

Khan, R., & De Angeli, A. (2007). Mapping the demographics of virtual humans. In *Proceedings of hci 2007 the 21st british hci group annual conference university of lancaster, uk 21* (pp. 1–4).

Kim, Y., & Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational technology research and development*, *54*(6), 569–596.

Lane, H. C., Hays, M. J., Core, M. G., & Auerbach, D. (2013). Learning intercultural communication skills with virtual humans: Feedback and fidelity. *Journal of Educational Psychology*, *105*(4), 1026.

Lane, H. C., Noren, D., Auerbach, D., Birch, M., & Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *International conference on artificial intelligence in education* (pp. 155–162).

Lee, C., & Lee, G. G. (2007). Emotion recognition for affective user interfaces using natural language dialogs. In *Ro-man 2007-the 16th ieee international symposium on robot and human interactive communication* (pp. 798–801).

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255–262.

McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, *31*(4), 1–11.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, *2*(1), 10–12.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100.

Nestel, D., & Tierney, T. (2007). Role-play for medical students learning about communication: guidelines for maximising benefits. *BMC medical education*, *7*(1), 1–9.

Ozogul, G., Johnson, A. M., Atkinson, R. K., & Reisslein, M. (2013). Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Computers & Education*, *67*, 36–50.

Riggio, R. E., & Lee, J. (2007). Emotional and interpersonal competencies and leader development. *Human Resource Management Review*, *17*(4), 418–426.

Salas, E., Wildman, J. L., & Piccolo, R. F. (2009). Using simulation-based training to enhance management education. *Academy of Management Learning & Education*, *8*(4), 559–573.

Schmid Mast, M., Kleinlogel, E. P., Tur, B., & Bachmann, M. (2018). The future of interpersonal skills development: Immersive virtual reality training with virtual humans. *Human Resource Development Quarterly*, *29*(2), 125–141.

Schrepp, M., Held, T., & LaugWitz, B. (2018). *User experience questionnaire.* Retrieved from `https://www.ueq-online.org/`

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? a meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1–39.

Schroeder, N. L., Romine, W. L., & Craig, S. D. (2017). Measuring pedagogical agent persona and the influence of agent persona on learning. *Computers & Education*, *109*, 176–186.

Sogunro, O. A. (2004). Efficacy of role-playing pedagogy in training leaders: some reflections. *Journal of management development*.

Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., ... Traum, D. (2013). Virtual humans for learning. *AI magazine*, *34*(4), 13–30.

Tan, S.-M., Liew, T. W., & Gan, C. L. (2020). Motivational virtual agent in e-learning: The roles of regulatory focus and message framing. *Information and Learning Sciences*.

Tcha-Tokey, K., Christmann, O., Loup-Escande, E., & Richir, S. (2016). Proposition and validation of a questionnaire to measure the user experience in immersive virtual environments.

Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human Behavior*, *27*(2), 741–749.

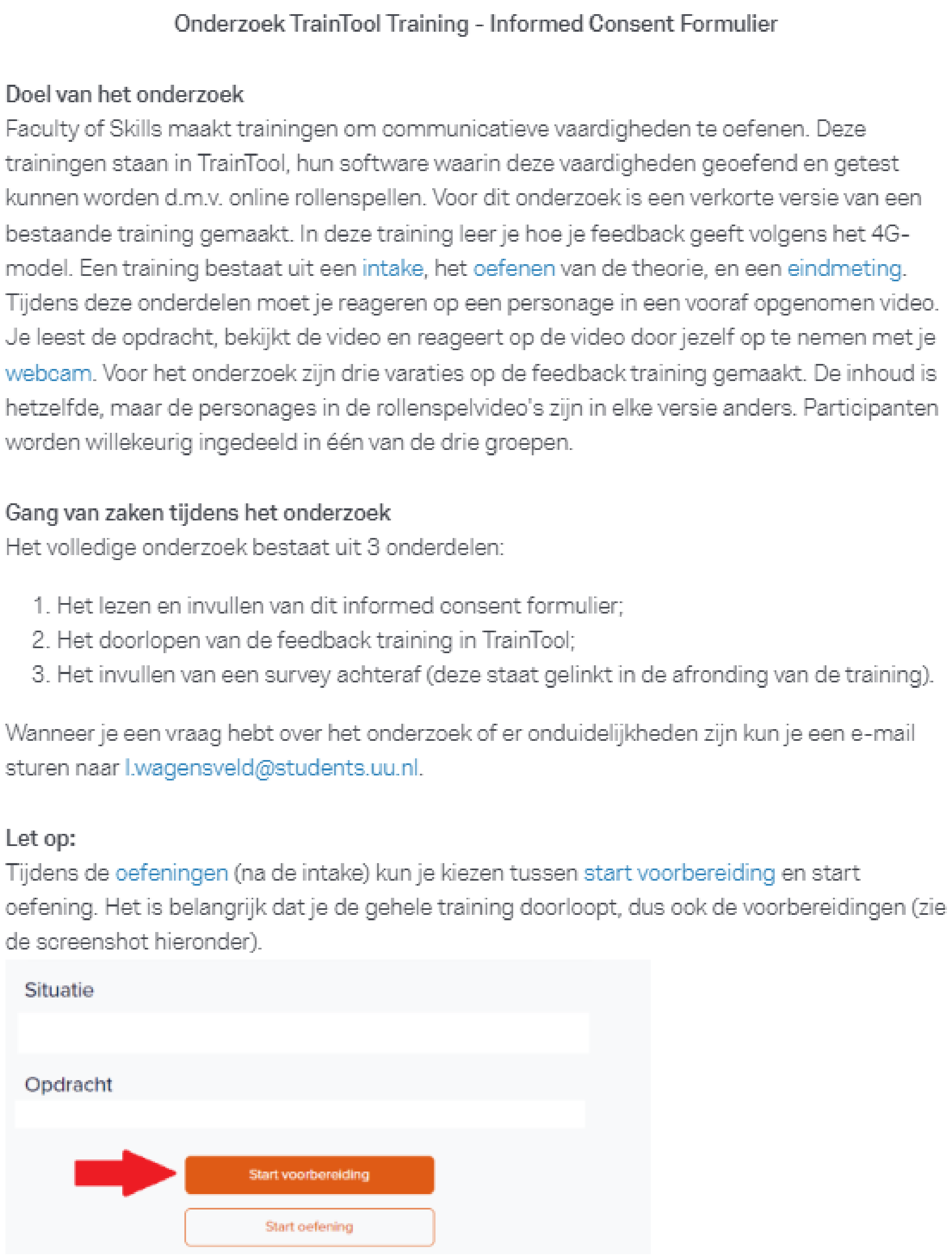
Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393–407.

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.*

Zipp, S. A., & Craig, S. D. (2019). The impact of a user's biases on interactions with virtual humans and learning during virtual emergency management training. *Educational Technology Research and Development*, *67*(6), 1385–1404.

A Appendix A: Informed Consent Form

The informed consent form, written in Dutch, can be found below.



Onderzoek TrainTool Training - Informed Consent Formulier

Doel van het onderzoek

Faculty of Skills maakt trainingen om communicatieve vaardigheden te oefenen. Deze trainingen staan in TrainTool, hun software waarin deze vaardigheden geoefend en getest kunnen worden d.m.v. online rollenspellen. Voor dit onderzoek is een verkorte versie van een bestaande training gemaakt. In deze training leer je hoe je feedback geeft volgens het 4G-model. Een training bestaat uit een intake, het oefenen van de theorie, en een eindmeting. Tijdens deze onderdelen moet je reageren op een personage in een vooraf opgenomen video. Je leest de opdracht, bekijkt de video en reageert op de video door jezelf op te nemen met je webcam. Voor het onderzoek zijn drie varaties op de feedback training gemaakt. De inhoud is hetzelfde, maar de personages in de rollenspelvideo's zijn in elke versie anders. Participanten worden willekeurig ingedeeld in één van de drie groepen.

Gang van zaken tijdens het onderzoek

Het volledige onderzoek bestaat uit 3 onderdelen:

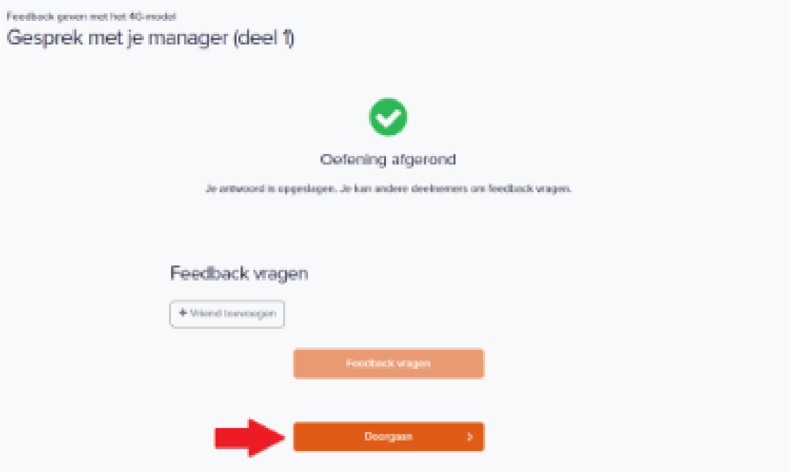
1. Het lezen en invullen van dit informed consent formulier;
2. Het doorlopen van de feedback training in TrainTool;
3. Het invullen van een survey achteraf (deze staat gelinkt in de afronding van de training).

Wanneer je een vraag hebt over het onderzoek of er onduidelijkheden zijn kun je een e-mail sturen naar l.wagensveld@students.uu.nl.

Let op:

Tijdens de oefeningen (na de intake) kun je kiezen tussen start voorbereiding en start oefening. Het is belangrijk dat je de gehele training doorloopt, dus ook de voorbereidingen (zie de screenshot hieronder).

Wanneer de oefening is afgerond krijg je de mogelijkheid om feedback te vragen aan vrienden, maar dit moet je overslaan. Klik hiervoor op doorgaan (zie de screenshot hieronder).



Feedback geven met het 4G-model

Gesprek met je manager (deel 1)

Oefening afgerond

Je antwoord is opgeslagen. Je kan andere deelnemers om feedback vragen.

Feedback vragen

Potentiële risico's en ongemakken

Tijdens de training moet je jezelf opnemen met je webcam. Dit kan in het begin een beetje onwennig of ongemakkelijk aanvoelen. Tijdens de intake en eindmeting is ervoor gekozen dat deelnemers zichzelf gaan beoordelen. Hierbij moet je een aantal stellingen aanvinken over wat je wel of niet hebt genoemd tijdens je opname. Hierbij wordt uitgegaan van de eerlijkheid van participanten, maar dit heeft ook als voordeel dat alleen jij de beelden te zien krijgt.

Vergoeding

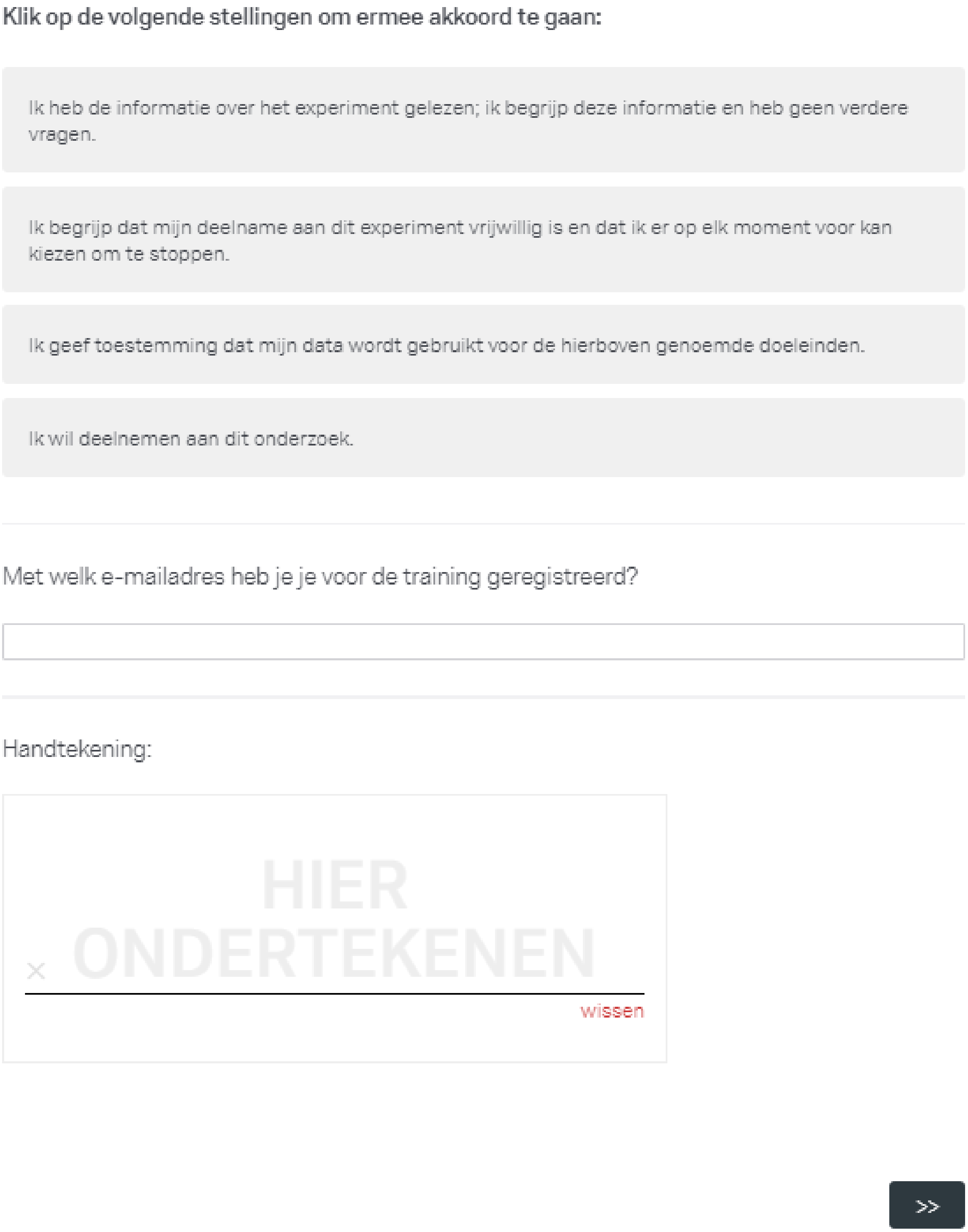
Omdat het onderzoek 20-30 minuten van je tijd kost kun je aan het einde van de survey aanspraak maken op een BOL.com bon t.w.v. €5,- als bedankje voor je tijd en moeite. Deze ontvang je wanneer het onderzoek is afgelopen (waarschijnlijk eind juli).

Vertrouwelijkheid van gegevens

De opnames die gemaakt zijn in TrainTool zullen niet gebruikt worden voor het onderzoek en worden 45 dagen na afloop van het onderzoek verwijderd. Antwoorden worden vertrouwelijk behandeld en digitale gegevens worden opgeslagen in een beveiligde omgeving. De resultaten van de intake, eindmeting en survey zullen geanonimiseerd worden. Publicaties die op dit onderzoek zijn gebaseerd, zullen geen informatie bevatten waarmee je geïdentificeerd kunt worden. De onderzoeksgegevens van het project kunnen worden bekeken en beoordeeld door departementen van Universiteit Utrecht.

Vrijwilligheid

Deelname aan dit onderzoek is geheel vrijwillig. Je kunt als participant jouw deelname aan het onderzoek op elk moment stoppen zonder hier een reden voor op te geven. Je hebt echter alleen recht op de vergoeding wanneer je het volledige onderzoek (training + survey) hebt afgemaakt.



Klik op de volgende stellingen om ermee akkoord te gaan:

Ik heb de informatie over het experiment gelezen; ik begrijp deze informatie en heb geen verdere vragen.

Ik begrijp dat mijn deelname aan dit experiment vrijwillig is en dat ik er op elk moment voor kan kiezen om te stoppen.

Ik geef toestemming dat mijn data wordt gebruikt voor de hierboven genoemde doeleinden.

Ik wil deelnemen aan dit onderzoek.

Met welk e-mailadres heb je je voor de training geregistreerd?

Handtekening:

B Appendix B: Survey questions

Below, an overview of the survey questions, translated from Dutch to English, is presented.

Demographic questions:

1. What is your email address?
   (Open text answer)
2. What is your age?
   (Open text answer)
3. What is your gender?
   (MC: male, female, non-binary, prefer not to say)
4. What is the highest level of education you have completed?
   (MC: primary school, secondary school, bachelor's degree, master's degree, doctorate degree, ...)
5. Which of the following best describes your current employment status?
   (MC: working full-time, part-time, student, unemployed, ...)
6. Was this your first time doing a training in TrainTool?
   (MC: yes/ no)

Questions regarding the training and the experience:

1. Please rate the following statements:
   (5-point Likert scale: strongly disagree (1) - strongly agree (5))
   (a) I thought the quality of the training was good.
   (b) I thought the quality of the characters was good.
   (c) It was easy to apply the theory.
   (d) I have learned a lot from the training.
2. Please rate the training based on the following criteria:
   (7-point Likert scale)
   "I thought the training was ..."
   (a) Annoying - Enjoyable
   (b) Complicated - Easy
   (c) Unpleasant - Pleasant
   (d) Not interesting - Interesting
   (e) Conventional - Inventive
   (f) Unattractive - Attractive
   (g) Bad - Good

   (h) Usual - Leading Edge

3. Please rate the following statements*:
   (10-point Likert scale: strongly disagree (1) - strongly agree (10))

   (a) My interactions with the role-play partners seemed natural. *(presence)*

   (b) The visual aspects of the role-play partners involved me. *(engagement)*

   (c) I was involved in the role-play experience. *(engagement)*

   (d) I felt stimulated by the role-play/ training environment. *(immersion)*

   (e) I became so involved in the training environment that I was not aware of things happening around me. *(immersion)*

4. What did you like about the training?
   (Open text answer)

5. What did you dislike about the training?
   (Open text answer)

** The category each statement belongs to is stated between the parentheses.*

Questions regarding character perception:
In this section of the survey, participants were shown images of characters they encountered during the training, followed by several statements about the characters. In total, participants were asked to rate four characters (two male, two female), using a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5).

1. I thought he/she was realistic.

2. I felt discomfort looking at him/her.

3. I thought he/she was credible.

4. I thought he/she was sympathetic.

5. His/her physical appearance looks appealing.

6. I felt engaged watching and listening to him/her.

7. His/her voice sounded natural.

8. His/her voice sounded pleasant.