HRI VR-Experiment RoboProfs

Fynn Aurand
Institute of Cognitive Science
University of Osnabrueck
Osnabrueck, Germany
faurand@uni-osnabrueck.de

Sebastian Beyrodt
Institute of Cognitive Science
University of Osnabrueck
Osnabrueck, Germany
sbeyrodt@uni-osnabrueck.de

Dominik T. Brockmann
Institute of Cognitive Science
University of Osnabrueck
Osnabrueck, Germany
dobrockmann@uni-osnabrueck.de

Johannes M. Dittrich
Institute of Cognitive Science
University of Osnabrueck
Osnabrueck, Germany
jdittrich@uni-osnabrueck.de

Yannik A. Heß
Institute of Cognitive Science
University of Osnabrueck
Osnabrueck, Germany
yhess@uni.osnabrueck.de

***Abstract*—The use of robots in educational settings is a promising research field. Prior research has shown that a robotic tutor can increase the enjoyment of learning as well as match the performance of human tutors. In our experiment, we examined the impact of positive and negative feedback from a robot teacher on the learning performance and learning behavior of participants in an educational setting. We created a memory retention task, in which the content of the lessons was presented by a robotic teacher in Virtual Reality. After every lesson, the participants had the option to repeat the lesson. The participants were split into two groups. One of the groups, the positive condition, received encouraging feedback from the robot, while the other group, the negative condition, received exhorting and discouraging feedback. The negative feedback was supposed to be especially unpleasant when the participant chose to retake a lesson. We tested two hypotheses: The first was the "performance hypothesis": Participants in the negative condition perform worse in the memory retention task, due to fewer repetitions and due to less motivation enforced by the negative robot teacher. The second was the "decreased repetition hypothesis": Participants in the negative condition repeat fewer lessons than those in the positive. The performance hypothesis was not supported by our experiment data, since the participants of the negative condition were slightly better in answering lesson-content questions. The decreased repetition hypothesis was supported by our experiment data since the participants of the positive condition repeated the lessons more often than the participants in the negative condition.**

***Index Terms*—educational robot, social robots, social-cues, emotional expressiveness, human-robot interaction, virtual reality**

I. Introduction

A. Idea of the Experiment

Going into this project, we had several ideas for what our experiment could look like. We were generally really interested in creating (strong) emotional reactions within our experiment and decided to pitch different ideas to each other. They ranged from having a virtual robot therapist, inspired by ELIZA [25], over examining the role of a robot teacher's emotional feedback during learning, to investigating the connection between a robot's design and the frustration as well as shame it could create. Ultimately, we decided on the second option and developed the idea further.

In particular, we were interested in what effect the nature of a virtual robot teacher's emotional feedback has on students. Some specific questions that we came up with were: Can a virtual robot elicit an emotional reaction in interactions with humans? Does such an emotional connection influence learning behavior? How does our knowledge about human-human interaction in learning situations translate to human-robot interaction? How could a virtual robot deemed to act as a tutor or lecturer look like? What components does it need to have to convey certain emotions?

We were fascinated how a robot could encourage and discourage a human in their knowledge acquisition process through emotional feedback on the learning process. As can be seen in the following documentation, the aspect of conveying strong but plausible emotions through ADA, as we named our virtual robot, was one of the main focuses of our work.

First, we will look at what research has already been done on teaching robots and socially interactive robots. In Section [II] we explain our experimental design and procedure. We will then give detailed documentation on the work we did in order to create the learning materials [III-A], an emotionally expressive robot [III-D] and to realize the experiment in Unity [III-J]. Section [IV] reports on our findings and Section [V] discusses the problems and impediments of our experiment and its design. Finally, Section [V-C] presents ideas on how to expand the project.

B. Prior research

Socially interactive robots are able to engage in social interaction and mainly used in domains in which the robot must exhibit peer-to-peer interaction skills. A socially interactive robot must send signals (e.g. social cues) to the user in order to provide feedback on its internal state. Such feedback can be provided through artificial emotions. Emotions play a significant role in human-to-human interaction, and the use of artificial emotions can help to facilitate believable human-robot interaction. Channels for emotional expression include vocalization, facial expressions, gestures, and body movement (see [12] and references therein).

One domain, where such a peer-to-peer interaction skill would be necessary, would be a learning scenario with an educational robot and a student. The main goal of a robot used in educational settings is to enhance the interaction between the student and the learning material by making the robot engage the student. Such socially engaging robots can be nearly as effective as human tutors [6] and can therefore be a cost-effective alternative that could be available to each student individually, especially if the robot is a virtual avatar. But in order to achieve this, the robot has to be socially interactive and needs to be able to express human equivalent behavioral cues of engagement. By using such cues, a relationship between the robot and the student can be fostered, which is able to increase the students' motivation to complete a certain task (see [6] and references therein). Hence, as described by Brown et al. in [6], much research has been conducted on implementing social characteristics in educational robots. We were especially interested in those studies that varied the type of feedback from the robot (positive, neutral, negative). They discovered, that students were more drawn to robots which gave them positive feedback. In [6], the researchers were able to show that a robot expressing a mixture of positive verbal and nonverbal cues can decrease boredom and increase enjoyment while doing an exam. They suggested repeating this experiment with a virtual avatar and comparing the results. We were very interested in doing this in a modified form, as we wondered if the effects described above that apply to physical robots are also found with respect to virtual robots. We combined our ideas into one experiment in which we wanted to research how the emotional feedback of a virtual teaching robot would affect subjects in their learning behavior and performance.

C. Aim of the Experiment & Hypotheses

The aim of the experiment documented in this paper is to find out if there are differences in the retention of the taught information (see Section [IV-C1] for the respective analysis) as well as whether either group would take the opportunity to repeat a lesson more often and whether the choice of the participants would change during the course of the experiment (see Section [IV-C2] for the respective analysis).

We hypothesized that the participants in the negative condition would presumably perform worse overall in the comprehension questions because we imagined that they would not only use less the chance to repeat a lesson in general to avoid at least some confrontation with the discouraging robot teacher but also retain less information independent of the repetition due to a lack of motivation.

Regarding the behavior to repeat a lesson, we assumed that the participants in the negative condition would repeat fewer lessons overall and also that their willingness to repeat a lesson would decrease during the course of the experiment.

II. Experiment Design & Procedure

1) Experiment Procedure & Components: Our experiment is made up of three parts. First, there is the introduction which is followed by the "lecture" in VR and, afterwards, the subjects should fill out a questionnaire. The questionnaire itself consists of three sections: Personal data, a quiz, called the "citizenship-test", and a general question for additional qualitative analysis (see Section [IV-D]) of the interaction. See Section [III-K] for more details on the questionnaire.

The lecture in VR was subdivided into 5 smaller lessons. The order of the lessons was randomized to avoid side-effects and all participants had the opportunity to repeat each lesson once directly after they had heard it the first time. See Section [III-A] for more details on the content of the lessons.

The participants were split into 2 groups, making up the two main conditions in our analysis of the collected data. One group received encouraging feedback from their tutor (i.e., from their robotic teacher) and the other group was confronted with discouraging feedback from the robot. The feedback differed depending on whether the subjects wanted to repeat the lesson or not but still remained negative when "not repeating" was chosen in the negative condition and positive when "repeating" was chosen in the positive condition.

Figure 5 shows a schematic representation of the experiment procedure with its three main parts – Introduction, VR experiment, and questionnaire. All 5 lessons were initially presented in a neutral manner. Afterwards, the subjects had the chance to get a retake of the lesson or directly move on to the next one. Based on the choice, the robot gave feedback of either encouraging or discouraging nature.

2) Conditions & Feedback: Every participant was either in the positive group or in the negative group aka the two conditions. They were designed to be polar opposites. Both conditions had a core message. The positive core message was: "you are doing well, you are smart, I positively affirm you. Whether or not you understand the lectures, you are still smart and doing well." The negative core message was: "you are doing bad, you are stupid, I dislike you and dislike having to work with you. Whether or not you understand the lectures, you are still stupid and doing badly."

The positive group received positive feedback all the time.[1] The negative group received neutral/daunting feedback when they did not request a repetition, but when they did, the robot would get aggressive and personal.[2]

The negative feedback had two purposes, firstly they needed to be a counterpart for the positive feedback in order to

[1] If they did not request a repetition ADA would say: "That's great! You did amazing! Not many subjects were able to understand it on the first try." If they did request a repetition: "Don't worry about it, most subjects need a second go."

[2] If they did not request a repetition ADA would say: "Would have been a pity if you didn't understand it. 95 percent of the subjects understood it right away. So don't become too comfortable and keep paying attention, you especially can't afford not to." If they did request a repetition: "Are you serious? You are the first subject that didn't understand this lecture. This lecture was not that difficult. I think you are not made for this. Why don't you stop wasting our time and just start to understand the damn lesson."

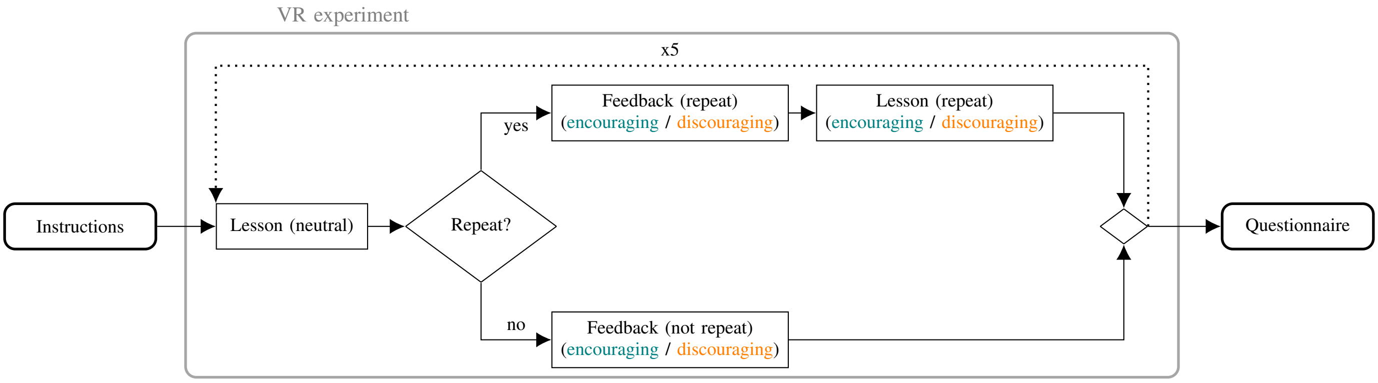


Fig. 1. Experiment Procedure

measure the effects on the performance. Secondly, the negative feedback was much worse, when the participants requested a repetition. Therefore, the participants had to decide whether to take the valuable repetition, but in turn, had to experience the mean and aggressive feedback or to just skip in order to avoid the aggressive feedback and miss out on the repetition.

III. A DETAILED DOCUMENTATION OF THE WORK

The order of the subsections of this chapter roughly reflects the chronological order in which we worked on or completed the things described in them during the semester. We started with the development of the robot design and learning content, which was also completed first.

A. Learning Content

For our lessons, we wanted to avoid that any participants could have prior knowledge. Therefore, we made up lessons about fictional aliens. The questions were about the creatures and the planets of that universe. In total, there were five lessons. One lesson about the planets, one lesson about the universe itself, one lesson about a specific planet (Hakol), one lesson about the species Prozu Falares and one lesson about the species Klono Pales. The following passage shows the script for the lesson about the universe: [3]

> *We are in a parallel universe, which entails 290,000 different species on 2,000 individual planets. The species with humanoid appearances are Zufa Rufu, Klono Pales, Jate Rales, Moka Nules and Prozu Falares. They all are intelligent creatures with feelings. Out of these species, Jate Rales and Zufa Rufu live peacefully on the planet Klux. The other intelligent species inhabit different planets.*

All lessons are about 100 words long so that it takes between 30-60 seconds to teach one lesson. In the lessons, we have facts about the world (numbers, superpowers, etc.) but there are also a lot of new names. We thought it could be helpful to visualize the lessons to make it easier to memorize them. We created the graphics ourselves via Figma [4] or used open commence license pictures mostly from Pixabay [5]. To get an overview: The creatures were created in Figma, the planets are from Pixabay. In addition to the images, we also added important facts to the presentation. So for example, the presentation not only contained a picture of a planet, but also the name of that planet. But in order to show these visualizations, ADA had to have a screen on which they could be displayed. With this, we already had our first requirement for her design.

B. Robot Design

Robots in educational settings can enhance the interaction between the content-user (the student) and the actual material content, therefore the design of the robot had to be carefully chosen [8]. Robots that have been used in educational settings before, did traditionally not have a full humanoid body [6]. In [8], Causo et al. reviewed the design of robots used in education and classified them into several categories. One of these was the “semi-humanoid” robots, which possess a head and a lower body and use wheels to move around.

1) Design of the lower body and head: We decided to model our robot ADA after such a “semi-humanoid” robot because it is believed that humanoid robots that are designed so that their appearance and movements are that of a human being have a negative impact on the student (see [6] and references therein). Using a “semi-humanoid” robot also allowed us to bring ADA to about eye level with the student by placing her on a table and didn’t require to give her a human-sized, probably intimidating, body. It also allowed us to integrate a 4:3 display into her lower body, on which she presents the lecture material. Since she is placed on a table, the student has a good viewing angle to look at the screen. Above the screen is a speaker from which ADA’s voice can be heard in the VR environment. Her actual head is connected to the lower body via a neck and is fully rotatable. The latter is important for her interaction capabilities and ability to express social cues (Section [III-D]). For the same reason, ADA also has

[3] All learning materials and scripts are uploaded to the GitHub Repository:https://github.com/roboprofs/documentation/tree/main/resources/stories

[4] https://www.figma.com

[5] https://pixabay.com

pivotable arms and ears that can also be lowered or raised. We also found that possible movement with wheels on the table is quite sufficient and well suited for working with e.g. proxemics. A more extensive description of her body can be found in Section [III-H1].

2) Design of the face: For the design of ADA's face, we heavily drew on the work of Kalegina et al. who searched for and analyzed robots with rendered faces. They found that many research robots have faces that are rendered onto a display. This gives complete flexibility over the design and allows to easily animate lip movements, blinking and facial expressions, which was very important to us with regard to the emotional expressiveness of ADA. Therefore, we decided that ADA should have a 19:6-Display as her head. They also discovered that most robots had a black face, a mouth, and eyebrows. With regards to the eyes, they figured out that most robots had eyes that took up around $\frac{1}{20}$ to $\frac{1}{10}$ of the screen space and were evenly spaced between the center and the edges of the screen. [13] Thus, we opted for a very simple yet emotionally expressive face consisting of a mouth and two eyes (occupying about 12.5% of the screen area and evenly distributed on the screen) and eyebrows on a black background. You can read more about the exact facial animations in Section [III-F].

C. Environment Design

We chose a classroom for the VR environment, as we felt that this was very suitable for a learning scenario. We kept the room as simple as possible to not distract from the robot. The key aspect of the environment was the table on which ADA is placed, as this determines her elevation, freedom of movement, and distance from the student. To create the classroom we used the "Low Poly Classroom" asset pack by Alberto Luviano[6] and the Archimesh Blender plugin[7]. [8]

D. Social Cues

Now that we had designed our robot and the environment in which she is situated. It was time to work on her emotional expressiveness, respectively her capability to convey social cues.

When creating the social cues, understandability and believability were the most important values. Prior research suggested, that robotic cues are generally well understood by human observers. Additionally, there was already plenty of research on the different channels of communication, such as: Speech [23], Facial Expression [4], Posture [9], Gesture [1], Gaze and Proximity [11], Color [5] as well as other social cues [17]. We collected these findings in a matrix, in which the columns represented the different channels of expressions. The channels of expression were called "Face", "Body", "Spatial", "Voice" and "Light". Those were further subdivided into basic parts such as e.g. "Position of Arms", "Eyebrow Position", "Eye Shape" etc. The rows represented the states of frames, meaning that each row represented a state of the robot, while the row below represented the next.[9]

There were five different scenarios in total. Scenario 1: The robot asks the subject whether or not he/she understood the lecture and if a repetition of the lecture is necessary. Scenarios 2 and 3 are only shown to the subjects in the positive condition. Which scenario is displayed depends on the subject's answer. If the subject reports not to need a replay of the lecture, scenario 2 will be shown. Otherwise, scenario 3 will be displayed. Scenario 2 shows the robot displaying happiness and excitement that the subject understood the lecture. It congratulates the subject and tells him/her that her performance was above average and positively reaffirms him/her. In scenario 3 the robot tries to divert attention away from the fact that the subject did not understand the lecture on the first try. It tries to cheer up the subject by pointing out the difficulty of the lecture. Scenario 4 and 5 mirror Scenario 2 and 3. Scenarios 4 and 5 are only shown to the subjects in the negative condition and which of the scenario shown depends on the answer of the subject.

Scenarios 4 and 5 were specifically designed to evoke a strong emotional reaction in the subjects. Therefore, we attempted to convey several underlying messages in the scenarios. Scenario 4, which is shown if the subject understands the lecture, starts with an expression of indifference towards the subject's "success". It continuous by slightly angrily urging not to let the "success" get to his head and to keep paying attention, because the success is not in any way special. Furthermore, the robot implies that the subject is not intellectually gifted, by telling the subject that they, in particular, can't afford not to pay attention. Scenario 5 was supposed to evoke the strongest emotional response. It combined a display of strong anger and frustration at the subject with reproachful underlying messages. The robotic feedback begins by ADA being annoyed at the subject for not understanding the seemingly easy lecture. The robots proclaim that the subject was the only subject not to understand the lecture. This annoyance quickly turns into anger, which makes the robot approach and berates the subject.

In order to make the robotic feedback believable, we had the robot reference objective standards rather than personal feelings. It was believed that a robot that would verbally reference his feelings e.g. "I am so happy with your success" or "I am getting annoyed" would not be believable because the robot is obviously not sentient. Therefore, we embraced the robotic nature and had the robot state "facts" by pointing out the ratio of success of other subjects, which we hoped, would be even more believable coming from a robot.

The Matrix was then evaluated by a human theater actor, who mimicked the instructions for the scenarios listed in the matrix as accurately as possible. In a second go, the actor played the scenarios in a way he found fitting (only having

[6] https://alberto-luviano.itch.io/lowpoly-classroom-pack

[7] https://docs.blender.org/manual/en/latest/addons/add_mesh/archimesh.html

[8] The final Blender model of the environment can be found in the GitHub Repository: https://github.com/roboprofs/unity/tree/main/Assets/Classroom

[9] The finalized and filled out version of the matrix can be seen here: https://github.com/roboprofs/documentation/blob/main/resources/social_cues/social_cues_matrix.xlsx

the text). The versions only differentiated slightly, this might have been due to the extensive research the matrix was based on or to the fact that the actor played according to the matrix before and was therefore strongly influenced by it. [10]

As Feldman [3] notes the way people imagine emotions to look like is not always identical to how emotions look in real life. Therefore, there seems to be a tension between understandability and authenticity. We choose to go for understandability because this was closer to the way the research was done. Additionally, we were aware that employing an actor who learned in a theater would also lead to a greater emphasis on understandability, by using exaggerated expressions.

After evaluating the performance by the actor, with a special emphasis on having the different channels of expression work together cohesively, we used the matrix as well as some of the footage of the actor as reference for the animation and the facial expressions. In the following sections, we will further elaborate on the technical process of creating and designing the different channels so that they convey certain emotions.

E. Speech

The expression of emotions is a fundamental feature of human-to-human interaction and it has been argued that emotion is “the major currency in which social interaction is transacted” (Zajonc, 1980, cited after [14]). Emotions may be described as brief and intense reactions to goal-relevant changes in the environment (Oatley & Jenkins, 1996, cited after [14]. There are many theories to describe emotions: For example, they can be described in terms of discrete categories (so-called basic emotions such as anger, happiness, sadness, or disgust). [7], [12] The discrete emotion approach primarily focuses on features that distinguish emotions from one another (Ekman, 1992, cited after [14]).

The features of emotions that are communicated with speech are either verbal (concerning the words) or nonverbal (concerning the features of the voice) [18]. Nonverbal signals are assumed to be particularly suitable for communicating emotive information. Amongst these nonverbal social cues, vocal cues are the most frequently reported. (see [14] and references therein). Common vocal parameters that govern the emotional content of speech are the voice pitch (level, range, contour), the speaking rate, the loudness, the number and duration of pauses. These are called prosodic features they are concerned with the rhythm/melody of the speech. There are also spectral parameters such as the articulation precision or the voice quality, that describe the sound of the speech [7].

The emotions diverge with respect to the combination and variation of these vocal parameters in the speaker’s voice [7] [14]. The listener uses these parameters to draw conclusions regarding the speaker’s emotional situation (see [2] and references therein). Listeners can recognize the vocal parameters of discrete emotions with an accuracy well above the chance level (see [14] and references therein). These vocal effects are consistent between speakers, with only minor differences (see [12] and references therein). Therefore, it is possible to generally obtain the values of these features for specific emotions by observing how a human’s voice changes according to the respective emotions [22].

For our experiment, we mainly used the four discrete emotions: anger, happiness, sadness, and disgust. Therefore, we searched the literature for some prosody rules for expressing these emotions. In the process we identified the following rules that were successfully employed to express these emotions:

1) *neutral*: Neutral or unemotional speech has a narrow pitch range compared to emotional speech, with the pitch tending to be normally distributed about the average pitch level. [2], [18]
2) *anger*: When we are angry, we tend to speak in a higher pitch. Also the pitch range becomes higher as well as the rate of pitch changes. We speak faster, enunciated with a strong high-frequency energy, tense articulation and a breathy voice quality. [18] [2] [20]
3) *happiness*: Speech uttered when happy, has an increase in pitch as well as a much wider pitch range. During pitch changes, there are smooth upward inflections. The overall speaking rate can be faster or slower and the articulation is normal. The voice quality is considered breathy. [2], [18], [20]
4) *sadness*: Sad speech has a slightly lower than average pitch with a narrow pitch range. During pitch changes there is a downwords inflection. The tempo is slow and the intensity low. We don not speak very rhythmically but with a lot of irregular pauses. The voice quality is resonant and the articulation slurred. [18], [20], [22]
5) *disgust*: The pitch is lower than average but the pitch range wider, when we are disgusted. There are a lot of downward inflections at phrase endings. Due to prolonged phonation time and increased pause length the overall speech rate is lower. the voice quality is grumbled and the articulation is normal. [18]

1) Emotionally expressive synthetic speech: Text-To-Speech (TTS) systems allow us to convert any input texts into a synthetic speech. The speech synthesizer translates the text into a verbal representation, which is then used (together with prosodic information) to produce the acoustic waveforms. [2] XML-based markup languages can be used to add prosodic information to a text in order to improve the outputted as speech samples [22]. Therefore, we were able to use the prosody rules that we identified above to incorporate the respective prosodic features annotate each response from ADA should using EmotionML and SSML. As a text-to-speech synthesizer, we used Azure Neural TTS, as it allowed us to fine-tune the text-to-speech output using the Speech Synthesis Markup Language.

The Speech Synthesis Markup Language (SSML) is an XML-based markup language for the generation of synthetic speech, with control over the prosodic features like pitch, contour, speaking rate and volume. The SSML tag `<prosody>`

[10] The resulting videos can be found under this link: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/reference_videos

is used for this purpose. Its attributes are:

- `<prosody pitch="value">`: This attribute indicates the baseline pitch for the contained text.
- `<prosody pitch="value">`: This attribute indicates the baseline pitch for the contained text.
- `<prosody contour="value">`: This attribute represents pitch contour of the contained text, used to describe the changes in pitch over.
- `<prosody rate="value">`: This attribute indicates the speaking rate for the contained text.
- `<prosody rate="value">`: This attribute indicates the volume level of the speaking voice for the contained text.

With these, we were able to annotate all of ADA's responses according to the prosody rules identified above.

When she was presenting the lectures for the first time, we simply used the default settings everywhere. When she repeated the lecture in the negative condition we used the following values: `rate="-2.00%"`, `volume="+10.00%"` and `pitch="-3.00%"`, while when she repeated it in the positive condition these were used: `rate="-2.00%"`, `volume="0.00%"` and `pitch="+2.00%"`.

However, we sometimes adjusted the speaking rate in such a way that no repetition in one condition took much longer than in the other, in order to prevent any advantages through this. We also made extensive use of the break tag `<break>` during the repetition in order to indicate important information.

For ADA's reactions to whether a subject wants to repeat a lecture or not we also had to deviate from the values suggested by the prosody rules above, because there the emotions were not always quite discrete and clearly distinguishable. Therefore, we additionally tried to adjust the values of the prosodic attributes so that the speech output sounded as similar as possible to the same text spoken by an actor.

In order to express discrete emotions, we also used the `<mstts:express-as>` tag. The voice "SaraNeural" that we used allowed the style "angry" that expresses an angry and annoyed tone, the style "cheerful" that expresses a positive and happy tone, and the style "sad" that expresses a sorrowful tone. This additionally added a change in voice quality and articulation to the contained texts. We used the speaking style "angry" for every repetition of the lecture in the negative condition and "cheerful" for every repetition of the lecture in the positive condition. For ADA's reactions, on whether to repeat a lecture or not, we also used different speaking styles whenever ADA was in a state that roughly coincided with this emotion and the chosen style contributed to the emotional expressiveness.

The Figure III-E1 shows a finished prosody-rich XML file. Overall, we generated 24 different audio files for ADA containing emotionally expressive speech.[11]

```
<s/><mstts:express-as style="angry">
    <prosody rate="-2.00%" volume="+10.00%"
    pitch="-3.00%" contour="(48%,+3%)(90%,-14%)">
        Well, good for you!
    </prosody>
    <prosody rate="-2.00%" volume="+10.00%"
    pitch="-3.00%">
        It would have been a shame, if you would not
        have been able to understand this lecture,
        because 95% of all subjects understand it
        right away.
        So don't become too comfortable, keep paying
        attention! especially you
    </prosody>
    <break strength="x-weak"/>
    <prosody rate="-2.00%" volume="+10.00%"
    pitch="-3.00%">
        can't afford to be distracted!
    </prosody>
</mstts:express-as><s/>
```

Fig. 2. Annotation of ADA's response when the subject does not want to repeat the lecture (in the negative condition), according to the prosody rules

F. Facial animations

A person's emotions and state of mind can be read from their face. Therefore, even the addition of simple facial animations contributes significantly to the liveliness and believability of a virtual avatar [24]. The primary facial components used are mouth (lips), cheeks, eyes, eyebrows and forehead. [12] Therefore, a characteristic speech in an animation not only depends on acoustic cues, but also on visual cues. Key components of realistic face animations are lip synchronization, blinking and facial expressions. [19]

However, Bennett and Sabanovic [4] provided evidence that a face only needs basic elements in order to be able to convey emotional expression. In their experiment they used a robotic consisting of 6 lines, two for the eyebrows, two for the eyes, one for the upper lip and one for the lower lip. Still, subjects were able to accurately identify the facial expressions. Given our robot design this seemed to be a fitting reference, since we did not want to do too many details in the face.

1) Facial expressions: In our experiment we used three actuators (eyes, eyebrows and mouth) to display seven facial expressions (neutral, admonishing, angry, disgusted, happy, sad and unimpressed). We came up with these 7 emotions when we watched the professional actor's videos, where he was recording ADA's lines. We filtered out the most distinctive and characteristic facial expressions. Based on this and the references from the basic emotion's literature [4], we then designed the eyes plus eyebrows and mouth in Affinity Designer [12]. This way we identified 7 eye states (neutral, admonishing, angry, disgusted, happy, sad and unimpressed) and 5 mouth states (neutral, angry, disgusted, happy and unimpressed) that can be combined to form these characteristic facial expressions. Overall, there are three facial expressions used for scenario 1-3 and four used for scenario 4 and 5 (two

[11] All 24 XML files used for the generation can be found here: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/speech. The 24 generated audio files can be found under this link: https://github.com/roboprofs/documentation/tree/main/resources/stories/audio

[12] https://affinity.serif.com/

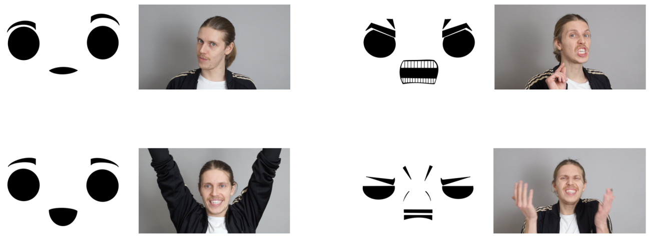


Fig. 3. Using the facial expressions of a professional actor as a reference for a first draft of the robotic expressions

each). [13]

Here it is important to add that these images are only shown on the displays when the mouth is in a rest position (for the images portraying the mouth in a certain emotional setting) and the eye is open (for the images portraying the eyes in a certain emotional setting). Otherwise, the standard images are used.

When ADA presents the lectures for the first time her eyes and her mouth are in a neutral state. When she repeats the lecture in negative condition, her eyes and mouth are in an angry condition, while when she repeats it in the positive condition they are in happy condition.

The facial expressions for ADA's reactions, to whether a subject wants to repeat a lecture or not, were especially interesting. When a subject chooses to repeat the lecture in a negative condition her first facial expressions is disgust (disgusted mouth + disgusted eyes) and then switches to angry (angry mouth + angry eyes). In the positive condition, however, she looks sad (neutral mouth + sad eyes). When a subject chooses to not repeat the lecture in a negative condition add reacts with an unimpressed face (unimpressed mouth + unimpressed eyes) and then switches to admonishing expressions (admonishing eyes + disgusted mouth).

2) Lip synchronization: Lip synchronization is the technical term for matching lip movements to the phonemes from an audio track. Phonemes are perceptually distinct units of sound that can distinguish one word from another [19]. The visual counterpart for a phoneme is called a viseme. A viseme represents the shape of the lips and position of the tongue when articulating an auditory syllable. Many phonemes however have ambiguous visual representations and map to the same viseme because they look the same on the speaker's face when they're produced [24]. Therefore, there is a many-to-one mapping between phonemes and visemes and each viseme depicts the key facial poses for a specific set of phonemes.

The visemes must be perfectly synchronized with the auditory phonemes in order to appear realistic. [19] This was another reason why we used Azure Neural TTS as a speech synthesizer. When using Azure Neural TTS, the speech output can be accompanied by a viseme ID and their offset timestamps. The viseme ID is an integer number that specifies a viseme. For English (US), there 22 different visemes, each depicting the mouth shape for a specific set of phonemes. [14] With the offset timestamps, we were able to construct the lip animations using a flipbook method. The flipbook method rapidly displays a list of consecutive visemes together with the auditory speech to create an impression of lip movement. (see [24] and references therein).

TABLE I
MAPPING BETWEEN THE VISEME SET USED BY AZURE NEURAL TTS AND THE SET OF VISEMES WE MADE

New Viseme ID	Old Viseme ID
A	1,6,11
B	4,5
C	2,3,8,9,10
D	7
E	12,13,14,15,16,19,20
F	18
G	14, 17
H	21
I	0

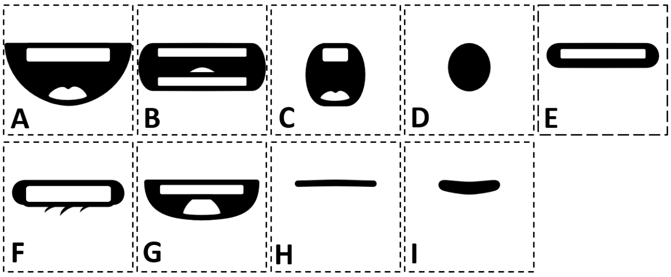


Fig. 4. Dialogue chart containing the 9 mouth shapes (visemes of our viseme set) that correspond to one or more phonemes (as shown in Table I). They were created in Affinity Designer

However, to make lip synchronization look realistic, it is necessary to construct these viseme sets carefully, since phonemes should be mapped to the correct visemes in order for the additional visual information to contribute to speech perception [24]. We now knew which phonemes belonged to one of the 22 visemes, but we had no reference points as for what the key facial poses were for them. Therefore, we decided to use a different set of visemes instead. We based our viseme set on the visemes in the Preston Blair (1194) phoneme series. The Preston Blair phoneme series (1194) is a popular set of visemes used in cartoons and research [24]. Our slightly changed two 2D adaptation of it can be seen in III-F2. We then proceeded to map the 22 visemes down to the 9 visemes we used as shown in Table I.

We then stored this information about the offset timestamps and the new Viseme IDs from A-I for every section of ADA's speech in an individual CSV file. There we also added the information about the emotional state of ADA during that

[13] The final designs for the eyes can be viewed here: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/eye/materials. The ones for the mouth here: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/mouth/materials

[14] The exact mapping can be found under the following link: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-speech-synthesis-viseme?pivots=programming-language-python#map-phonemes-to-visemes

response according to what was described in the Section [III-F1]. [15]

3) Blinking behavior: Blinking behavior is an important part of human nonverbal communication. We based our blinking patterns on the work of Lehmann et al. In [16], they describe a method they used to model human-like robot blinking. They divided each blink into three phases, the attack phase when the eye is closing (mean length = 111ms, standard deviation = 31ms), the sustain phase when the eye is closed (mean length = 20ms, standard deviation = 5ms), and the decay phase when the eye is opening again (mean length = 300ms, standard deviation = 23ms). The mean average time until a blink occurs is 5200ms with a standard deviation of 3700ms. Additionally, there was a 15% chance that the robot would blink again after the first blink. [16]

We have taken these values and adjusted them to work for 2D eyes on a display. Approximately every 5.2 seconds (normally distributed) ADA's eyes change from the image of an open eye to that of a half-open eye. This then changes to a fully closed eye (attack phase), which is then visible for a moment and then changes again to a half-closed eye (sustain phase), which then again opens fully (decay phase).

In this manner, we created a blinking behavior of equal length for each section of ADA's text. We then stored this information about the offset timestamps and the state of the eye (closed, half closed or open) for each section in a CSV file. There we also added the information about the emotional state of ADA during that section according to what was described in the Section [III-F1]. [16]

G. Movement, Gesture, Posture, Proximity and Color

It was clear from the start, that would not be able to replicate human like behavior. Luckily this was not necessary, the work of Rooij et. al. [21] as well as Emgen et. al. [10] successfully used abstraction to overcome the technical limits of today's robots. The idea behind abstraction is to only focus on the essential features necessary to convey the intended message. [21] Emgen et al. provided evidence that these abstracted channels alone can be so powerful, that participants can identify emotions even tho their robot did not have facial expression. In their study, they used for example body posture (upright for happiness, lowered for disappointment, towards the viewer for curiosity etc.) and ear positioning (e.g., flattening for fear and shame, lifting for happiness). [10] Other researchers like Bethel & Murphy [5] also included body movements, orientation and color. They associate fast and expansive movements with happiness, while slow and hesitating movements are associated with depression. Also, the orientation can convey information about the attentiveness of the robot. Facing towards the user tends to give the impression that the robot is paying attention to the user, while being faced sideways to the user is associated with the opposite. This also influences the effects of proximity. Low proximity is associated with intimacy. Unwanted entering of the intimate zone by a robot, especially when it enters head on, is perceived as threatening. Furthermore, colors can convey emotional states. It should be noted that the emotions associated with certain colors are not always clear. Red for example is associated with anger and with affection. So, we tried to make sure, that colors were only used in contexts where their meaning seemed clear and definite. In the positive condition, we used white, yellow, green and blue. These are associated with calmness, cheerfulness, pleasantness and joy. In the negative condition we used no light (black), orange and red, which are associated with disturbance, hostility and anger. [5]

H. Robot in Blender

Now, that we knew what social cues we wanted ADA to convey, we also knew the precise requirements our robot design had to meet. We felt that the design we had created in Section [III-B] was still suitable and started to create a Blender model of ADA.

1) Appearance: The model of ADA consists of a main body to which the wheels, arms and neck are attached to. The arms are in a simple rectangular shape and connected to the body using a rail system. The head is in rectangular shape as well and sits on top of a ball joint attached to the body with a tube-shaped neck. On each side of the head is an ear, a small plate connected using a similar rail system as the arms. The entire model uses rounded corners and edges in order to reach a more natural aesthetic by using a bevel modifier.

The textures are held simple to not distract from the important elements of the robot. In total the model of ADA uses four different textures. The body and the head use a metallic texture that has a pattern on it to not look too clean. A darker metallic texture sets the wheels visually apart from the body that they are attached to. The ears and the border of the screen on the body are in a matt dark grey tone. By giving them a different texture, they stand out more, especially the ears that are rather small but play a role in expressing emotions get more attention. The speaker, that is attached to the body above the screen, has a fabric pattern. All textures are public domain materials taken from ambientCG. [17]

2) Armature: The armature of our robot model consists of six bones that define the movements that are possible for ADA to make. The first bone is the root bone, it specifies and restricts the position and rotation of the model in global space. ADA has to be able to move forward on the table to meet the requirements of multiple social cues. Therefore, the root bone allows free movement on the x and y axis but not the z axis which would be an upwards/downwards movement. Rotation is possible on the z axis only, aligned with ADA's ability to drive freely on the table.

[15] All 24 CSV files used for the lip animation during ADA's response can be found in the GitHub Repository: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/mouth/data

[16] All 25 CSV files specifying ADA's blinking behavior can be found in the GitHub Repository: https://github.com/roboprofs/documentation/tree/main/resources/social_cues/eye/data

[17] The final blender model of ADA can be found under the following link: https://github.com/roboprofs/documentation/tree/main/blender

TABLE II
BONE CONSTRAINTS
(- NO MOVEMENT, * NO CONSTRAINTS)

	rotation			*translation*		
bone	**x**	**y**	**z**	**x**	**y**	**z**
root	-	-	*	*	*	-
neck	-3° to 182°	-12° to 12°	*	-	-	-
ears	-	-	-	-	-	path
arms	-	*	-	-	-	path

To perform the social cues and specially to emphasize certain words, gesticulation using arms is very important. It is controlled by two separate bones, one for each arm. The rotation is restricted on the local y axis which is the rotation that is done when pointing at something in front of you. Regarding translation, the arms can only be moved up and down and only by a small amount. This restriction is realized by using a hidden path that sits on top of a metal bar attached to the body of the robot. Attached to the arm is a small piece that sits inside the bar. The whole system functions as a rail, the arm bone can only be moved along the hidden path on top of the bar. Lowering the arms can for example mimic a dropping of the shoulders used in various social cues.

ADA's head sits at a ball joint on top of a tube representing the neck. The head can be rotated on the y axis to make a tilting of the head to the sides possible. Further, it is crucial for the head to be rotatable on the x axis for ADA to look down/up which is especially important to be able to look at the participant when moving towards them. Additionally, no limits on the z axis allow a shaking of the head and to look to the sides. These actions are substantial for the social cues, for example tilting the head to a side can show curiosity, shaking the head frustration and looking down to the participant while being in front of them seems reproachful. The limits of the rotation on the x and y axis are in order to stop the head from being able to move through the neck.

Though not the first body part one would think about for expressing emotions, ears play a significant role in inducing reactions from social cues as well. For this reason, the decision was made that ADA's design should include at least a fairly simple representation of them. The ears are not rotatable but can be moved along a rail using the same mechanism the arms use as described above. This is already enough to support the expression of emotions like curiosity and disappointment.

Table II shows the rotation and translation constraints of the bones in the robot model as defined in Blender. A minus sign corresponds to fully restricted motion along the respective axis, a star ("*") denotes full freedom in movement and "path" describes a "clamp to" constraint using a hidden path.

I. Animations

1) Gestural Behavior: After the completion of the robot model and the elaboration of the social cues matrix, we could now start to create animations. The goal for the animations was to successfully elicit full-featured emotional reactions in the participants. While we extensively used the social cue matrix (see Section [III-D]) that we compiled using scientific literature, we also animated ADA to match the reference videos by a professional actor. This helped a lot in making out the key moments for a convincing emotional animation.

Those key moments are the first keyframes set in the process of animating. Then more and more keyframes are added in between to obtain a fluent and natural motion of the various body parts of ADA. For the keyframes, we used the audio files to match the gestures with the speech. It allowed us to emphasize certain words or phrases in a perfectly timed manner.

In addition to the animations for every different scenario, we used looped animations for the lesson. Standing still would have been perceived as unnatural but a specially made animation for every second of every lesson is not necessary either because the important social cues are done in different scenarios.

2) Blinking & Talking Behavior: The face expressions and movements were constructed from 3 components. As described above (see Section [III-F3] and Section [III-F2]), *csv-files* – 25 for the blinking behavior and 24 for the visemes – for each animation sequence were created determining the state of the mouth and the eyes respectively at different times during the animation. Those states correspond to different *materials* for the eyes and 13 materials fort the variations of the mouth shape. Together with the 24 audios (see Section [III-E1], the materials and csv-files were integrated and synchronized in Unity.

Because we used a more complex shader in Blender (that was thankfully already freely available online [18]) to obtain a pixelized style for the display representing the face of ADA, we could not simply import the materials for the mouth and eye states into Unity, but instead had to bake them as new, individual textures first. Due to the fact that we baked those textures as simple black and white images, we were, furthermore, able to later on control the color and glow of the displayed pixel images in Unity. This was important as the color was another important social cue we integrated into our experiment (see Section [III-G]).

With all these components, we were then able to program scripts to read in the csv-file, on which bases we would – in the respective scenarios and predefined times – switch the material of the display used as ADA's face. The display consisted of three segments, so that we could change the material of left eye, mouth, and right eye independently, which saved us a lot of work. Depending on the current condition we also changed the colors of the white pixels of the materials in the script.

J. Implementation in Unity

The implementation of the experiment procedure in Unity used the separation of the different scenarios in 11 stories. Stories 1-5 denote the five lessons. These stories exist in three different variations: neutral, negative, and positive (the latter

[18] https://blenderartists.org/t/animatable-pixel-matrix-display/1126033

two for the repetition of the lessons). Story 6 is the question whether the participant wants to repeat or see the next lesson and 7/8 the corresponding reaction from ADA, depending on the answer. The stories 9 and 10 are introduction and ending respectively. Story 11 is the transition between two lessons.

The implementation makes use of a Controller class with multiple manager classes each in charge of a different aspect of the execution of a story. AudioManager starts and stops the audio clips of ADA's speeches that are saved in Sounds. The class FaceManager manages the change of the eyes and the mouth movement that both match the current story and ADA's talk. Data necessary for these animations is read in using CSVReader and files in CSV format for every story in every mood (see Section [III-I2]). The videos of the lessons that are displayed on the screen on ADA's body are played by VideoManager. Starting the correct animations is done by the Controller and an Animation Controller holding the animation of each story in every mood possible. For the logging of data, the class Logger is in charge. It saves triplets of time, an event type and a value in one log file per participant (see Section [III-J2]).

The right sequence of pairs of stories and the matching mood is implemented by a combination of the Controller and Trial class. A Trial is the execution of one lecture. It includes the stories of the lesson itself and the question whether the participant wants to repeat the lesson. By receiving the input from the Controller given by the participant using the VR controllers (see Section[III-J1]), the class manages whether the lesson is played a second time or this trial is finished. TrialSpecial is used for the stories 9, 10 and 11 which are the introduction, ending and transition between two lessons respectively. The reason for a special class for these stories only, is the fact that they do not occur in a predefined, repeating manner rather than being at the start, end or in between trials. The sequence of trials is implemented using a stack that is filled at the beginning of a session. Each time a trial is finished it gets removed from the stack until the stack is empty. The interaction between a Controller and Trial class allowed us to have an organized control of the sequence of actions and the overall experiment procedure. The full Unity implementation can be found in our GitHub repository[19].

1) Subject Controls: We decided to keep the controls as simple as possible to make the interaction the most intuitive. The participants have a controller in each hand and only the trigger buttons are relevant for this experiment. The question whether the participant wants to repeat the heard lesson or rather wants to keep going with the next lesson can be answered using the controllers. An input using the left controller repeats, using the right controller skips the repetition.

For the implementation of the interaction, we used SteamVR. The unity plugin handles the inputs from the VR controllers.

2) Data Logging: In addition to the questionnaire, the data logging from within the VR experiment is an essential part of the data gathering. The data provides the foundation to test our hypotheses.

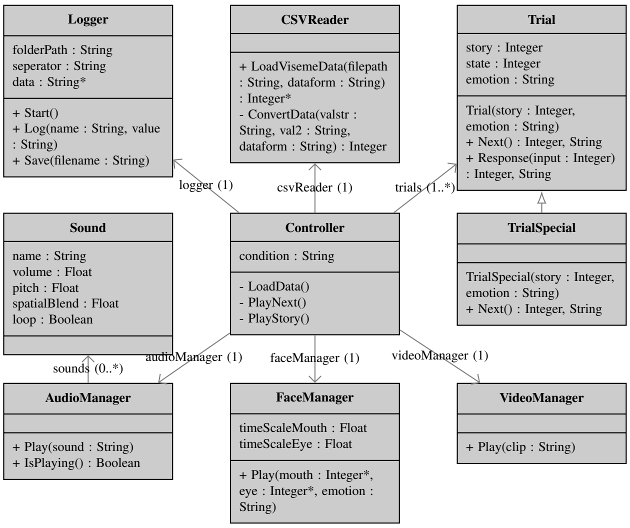


Fig. 5. UML Diagram

All data is logged as "events" with information about when they occur – measured in milliseconds since experiment start – and which values they take on. There are two distinct types of events that differ in the number of times they get recorded in one session: single-occurrence events and multi-occurrence events. The single-occurrence events encompass the events "condition", "lessons", and "participant_id". They are set in the beginning of each session and remain immutal thereafter. The multi-occurrence events comprise the events "button_press", "hmd_pitch", "repeat_lesson", and "starting_story".

Each participant has to be identifiable using a unique ID. They allow matching the data from the questionnaire to the data gathered during the experiment. Furthermore, the unique identifier must provide anonymity for the participants, which means that it should not be possible to associate any data to a participant. To provide both, uniqueness and anonymity, we decided to make use of Universally Unique Identifiers (UUIDs) of version 4 as defined in the IEEE standard RFC 4122 [15].

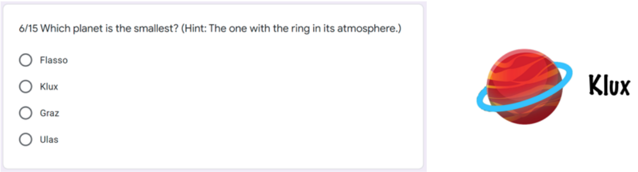
The event "condition" specifies under which condition the experiment is conducted. The order in which the lessons are held is logged using numbers from one to five representing the lessons. Each participant hears every lesson but in a randomized order.

On the press of one of the two relevant buttons, the event "button_press" is logged along with the information which controller is pressed, the one in the left or right hand. Additionally, if the press is the answer to the question whether the participant wants to repeat the lesson, the event "repeat_lesson" is recorded with the according truth value (true for repetition, false otherwise). This is the most relevant information because it allows us to test the decreased repetition hypothesis.

[19] https://github.com/roboprofs/unity

TABLE III
LOGGING DATA STRUCTURE
(POSSIBLE VALUES GIVEN IN REGEX)

<table>
<tr><th>event</th><th>possible values</th></tr>
<tr><td><i>button_press</i></td><td>(left;repeat|right;next)</td></tr>
<tr><td><i>condition</i></td><td>(negative|positive)</td></tr>
<tr><td><i>hmd_pitch</i></td><td>([1-2]?[0-9]?[0-9]|3[0-5][0-9]|360)\.[0-9]{4,5}</td></tr>
<tr><td><i>lessons</i></td><td>([1-5];){5}</td></tr>
<tr><td><i>participant_id</i></td><td>[0-9a-f]{8}-([0-9a-f]{4}-){3}[0-9a-f]{8}</td></tr>
<tr><td><i>repeat_lesson</i></td><td>(True|False)</td></tr>
<tr><td><i>starting_story</i></td><td>[1-5](neutral|negative|positive|(6|9|10)neutral|(7|8|11)(negative|positive)</td></tr>
</table>



6/15 Which planet is the smallest? (Hint: The one with the ring in its atmosphere.)

○ Flasso
○ Klux
○ Graz
○ Ulas

Fig. 6. Example screenshot from the presentation and the question in the questionnaire (nominal scale)

The event “starting_story” records the starting of a certain story in a condition (neutral, negative or positive). Logging the exact starting timestamp of a story helps to match interesting reactions of the participants to a situation. To observe these reactions, the pitch of the VR headset in degrees is noted every second with an accuracy of 10^{-4}.

Table III shows the different events with corresponding possible values written in the standard regular expression notation.

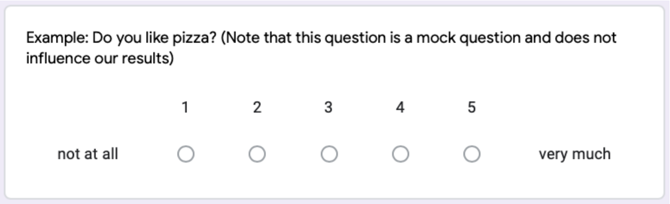
K. Questionnaire

Now, the last thing that needed to be done was to construct a questionnaire before conducting the experiment. We worked with “Google Forms” to construct our questionnaire, which was easy to use but also very effective. [20]

> *Do subjects feel uncomfortable to say that they did not understand a topic differently depending on the immediate feedback of a Robot teacher?*

For that, we decided to construct a questionnaire with mainly quantitative questions. The questionnaire was divided into three parts. The first part was about the learning content (Section [III-A]) of the lessons from the robot. We used a nominal scale for that - meaning that we had multiple-choice questions. That was important so that we could evaluate the answers fast and objectively.

Most questions of the first part that regarded the lesson’s content were about facts/statements that were also stated on the presentation that was displayed on ADA’s body. (See Figure 6)



Example: Do you like pizza? (Note that this question is a mock question and does not influence our results)

	1	2	3	4	5	
not at all	○	○	○	○	○	very much

Fig. 7. Example question that should explain the question format (ordinal scale)

The second part was about the human-robot interaction. It is hard to measure feelings in a questionnaire, so we decided to use an ordinal scale with 5 different answer options (from 0 = “not at all” to 5 = “very much”) to avoid black and white thinking. In this part of the questionnaire we also added one qualitative question, in which the participants were asked to write down their impression of the robots appearance.

The third part consisted of demographic questions and one question regarding feedback (“Do you have feedback?”).

In total, we had 32 questions. The questions were distributed in the following way: 16 content-related questions, 12 interaction-related questions, 3 demographic questions, and one feedback question.

However, one of the content-related questions and one of the interaction-related questions were example questions that should explain the question format. Before the example questions, we added short introductions to the question format, so that the participants got an overview of their task. (See Figure 7)

IV. RESULTS

All the data that was collected and will be analyzed in this Section, can be found together with the Rmd files of the analysis, in the GitHub Repository of this project[21].

A. Descriptive Statistics

Our experiment was conducted with 20 participants of which 10 identify as female and 10 as male. The participants were randomly assigned a condition with an equal number of participants in both conditions.

Figure 9 and Figure 8 provide an overview over the collected data. They show the proportion of correct answers in the “citizenship test” (comprehension questions) for each participant and the proportion of correct answers for each question respectively.

The subjects repeated about 1.053 ($\approx 26.3\%$) of the lessons 2 to 5 on average and the mean of the performance in the comprehension questions was 51.7%. Before cleaning the data, the lowest score achieved was 20% (before excluding data). The highest achieved score was 80% (before excluding data). For the respective values for each condition individually, see Table IV.

Figure 9 and Figure 8 also show the threshold set as the exclusion criterion which is further discussed in Section [IV-B].

[20] The whole questionnaire can be found here: https://github.com/roboprofs/documentation/blob/main/resources/stories/questionnaire.pdf

[21] https://github.com/roboprofs/documentation/tree/main/analysis

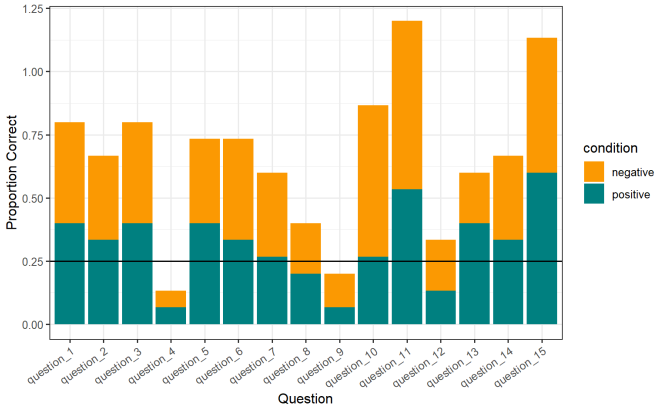


Fig. 8. Overview over the proportion of correct answer from the subjects in each comprehension questions (horizontal line denotes inclusion threshold)

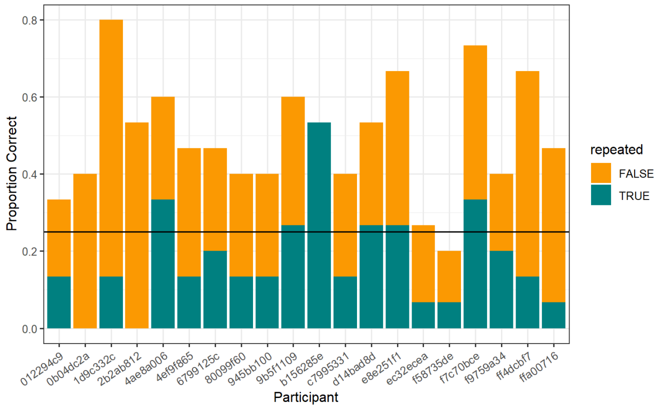


Fig. 9. Overview over the proportion of correct answer in the comprehension questions per participant (horizontal line denotes inclusion threshold)

B. *Tidying & Preprocessing*

To be able to use the data that we collected during the conduction of our experiment for analysis, the data had to be combined first. As described in Section [III-J2], we gathered data from within the VR environment and from the questionnaire afterwards independently but made sure to use a UUID – i.e. a unique, but random identifier – to allow us to anonymously join the respective data during the preprocessing step of our data analysis. Additionally, Unity logfile names of test runs were marked with a '+' sign in the beginning for dropping them easily from the data to be analyzed. We also added information about whether each answer of the "citizenship test" (comprehension questions) was correct or not and formatted the data in a (to a certain degree) *tidy* [26] way.

TABLE IV
PROPORTIONS OF REPETITIONS & CORRECT MEMORY RETENTION
(PER CONDITION, ROUNDED TO THE 3^{rd} DECIMAL PLACE)

	repetition			*performance*		
condition	**mean**	**median**	**sd**	**mean**	**median**	**sd**
negative	0.225	0.25	0.219	0.533	0.542	0.137
positive	0.306	0.25	0.208	0.5	0.5	0.195

The logfiles from within the VR experiment had to be processed more extensively as we had logged the data in a very general manner by only using the headers 'time', 'event', and 'value' and storing everything (participant id, condition, order of lessons, button presses, start of stories, pitch) in it. For more on the tidying and preprocessing please refer to the Rmd-file.[22]

We decided to both exclude comprehension questions that were answered correctly by less than 25% of the participants and participants that answered less than 25% of the comprehension questions correctly. As all the questions had four answer options to choose from 25% would have been the value that should have been correct if the comprehension question would have been answered by chance. Therefore, *question 4* and *question 9* were excluded as well as 1 participant. We assume there to have been problems in comprehension (for example due to unsuited formulation of the questions and not sufficient knowledge of the English language respectively) which is why the data seems to be not relevant for the investigation of our hypotheses.

Furthermore, we had to remove the first trial for all participants, as most of them were confused by ADA asking two different questions directly after another: "Did you understand everything, or should I repeat the lecture?". Some answered the first question, however, the second question – as also shown on ADA's display – was actually the relevant one. As the two questions were exactly the opposite concerning their answer, there were a few participants for whom the given answer was not the one they had intended.

C. *Quantitative Analysis*

To repeat again: With our experiment, we were interested in testing differences in the retention of the taught information (in the following called the 'performance hypothesis') where we hypothesized that subjects in the negative condition would perform worse both overall due to fewer repetitions and independent of repetitions due to less motivation enforced by the negative robot teacher. And, in addition, we wanted to know whether either group would take more often the opportunity to repeat a lesson and whether the choice of the participants would change during the course of the experiment (in the following called the 'decreased repetition hypothesis'). Our thoughts on this were that participants in the negative condition would choose less often to repeat a lesson than those in the positive and, moreover, would be less willing to repeat a lesson the further into the experiment they were.

1) Performance Hypothesis: By means of Bayesian regression modeling and hypothesis testing, we analyzed the distributions of the memory retention scores for both the

[22] The R-Code was uploaded to the GitHub Repository: https://github.com/roboprofs/documentation/tree/main/analysis/preprocessing.Rmd, https://github.com/roboprofs/documentation/tree/main/analysis/quantitative_analysis.Rmd

negative and positive condition and investigated whether it is reasonable to assume that participants in the negative condition performed generally worse than those of the positive condition. Coming up with a very low evidence ratio of clearly under 1 (posterior probability around 0.40), we had to reject our hypothesis. As a matter of fact, the average performance of both groups was extremely similar only differing by about 0.033 with a slight tendency towards a better performance by the group with condition negative.

For analyzing the performance depending on the different conditions – "negative vs positive feedback" as the main condition of interest and "repeated vs not repeated" as an additional condition – we made use of contrast coding and Bayesian regression modeling. By doing so we were able to examine whether there was a "main effect feedback", a "main effect repeated", and whether both conditions interacted with each other. Using the programming language `R` and the `brms` package, we created a Bayesian regression model explaining the proportion of correct answers by the multiple predictors "repeated VS not repeated" and "negative condition VS positive condition".

Testing for a "main effect repetition" – i.e. whether repetition alone would have a significant influence on the proportion of correct answer – with this model yielded with an evidence ratio of under 0.1 (posterior probability around 0.05) surprisingly no evidence at all that subjects that repeated a lecture performed better compared to those that did not. Testing for a "main effect feedback" – i.e. whether the type of feedback (encouraging or discouraging) alone would have a significant influence on the proportion of correct answers – with this model yielded an evidence ratio of under 2.1 (posterior probability around 0.66) suggesting no evidence that subjects in the positive condition performed better compared to those in the negative contradicting our initial hypothesis.

A possible explanation could be that the participants in the negative condition might have reacted with an act of defiance or by wanting to appeal to ADA and trying harder to understand the content (to understand why this might be a reasonable explanation see Section [IV-D], analysis of Q4.5).

Testing for interaction between repetition and condition yielded fairly similar results: With an evidence ratio of clearly under 1 (posterior probability around 0.46), there is no evidence that the interaction between whether participants repeated a lesson or not and in which condition they were influenced the proportion of correct answers in the "citizenship test" afterwards.

2) Decreased Repetition Hypothesis: Looking at the overall mean of repetitions in the negative versus the positive condition – using again Bayesian regression model and hypothesis testing – the data shows strong evidence (evidence ratio just above 10 and posterior probability of around 0.92) for the hypothesis that subjects in the negative conditions were generally less willing to repeat lessons than subjects in the positive condition.

For testing our decreased repetition hypothesis, we proceeded similarly as with the performance hypothesis. However, as we were only interested in whether the willingness to repeat a lesson decreased during the course of the experiment – i.e. we are only looking at the number of the trial as a single predictor – our Bayesian regression model for this hypothesis was slightly simpler.

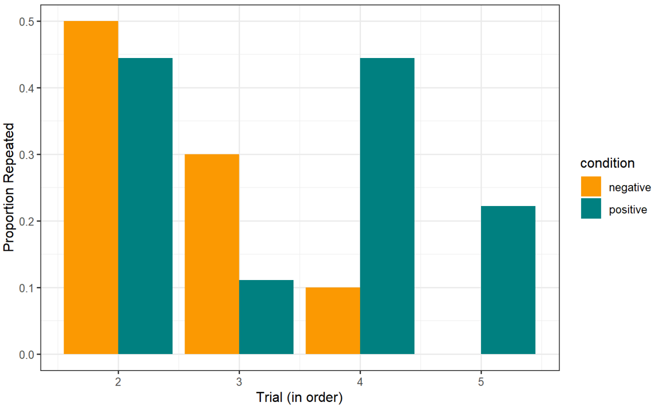


Fig. 10. Comparison of the proportion of participants in both conditions that took the repetition (note: per trial, not per lessons as the order varied)

Looking at the decrease in the positive condition first, we found strong evidence for a decrease from trial 2 to 3 (remember: trial 1 was excluded from the analysis due to confusion) with an evidence ratio over 200 (posterior probability of over 0.99). Contrary, from trial 3 to 4 the willingness to repeat a lesson drastically increased again to about the same value as in trial 2 resulting in an extremely low evidence ratio of clearly under 0.01 (posterior probability under 0.007) for the decrease hypothesis. From trial 4 to 5 the decrease is supported again with an evidence ratio of above 20 (posterior probability of over 0.96) for the positive condition. Additionally, trial 3 showed the lowest willingness of subjects for a repetition in the positive condition.

With evidence ratios of above 38, 40, and 5.2 (posterior probabilities of around 0.98, 0.978, and 0.84) – for a decrease from 2 to 3, 3 to 4, and 4 to 5 respectively – we found strong evidence supporting our hypothesis of a consistent decrease for the negative condition. In fact, none of the subjects in the negative condition repeated the last the lesson in the last trial – which also explains the much lower third evidence ratio compared to the first two. This steady and consistent decrease can also clearly be seen in Figure 10.

3) Pitch Analysis: We analyzed the recorded pitch data for moments that show sudden or large head movements made by multiple participants in the same stories and at the same time. Additionally, we wanted to find out whether there are general differences between the two conditions, e.g. more movement overall in one of the conditions compared to the other. Unfortunately, we could not find interesting reactions or significant differences regarding the recorded head pitch data.[23]

[23] https://github.com/roboprofs/documentation/blob/main/analysis/pitch_analysis.Rmd

D. Qualitative Analysis

The first question Q1 ("1/10 Did you understand the content of the lectures?") and the third question Q3 ("3/10 Did you try your best at understanding the lectures and answering the questions?") were used as requirements for a participant's data to be included and together with question 2 Q2 ("2/10 Did you find the lectures to be too difficult?") as general feedback of the difficulty. If someone would have answered question 1 or question 3 with 1 = not at all, this participant's data would have been eliminated. No participant gave the answer 1 for any of the questions. The results for Q1 (mu = 4.2, M = 4 and sd = 0.76), Q2 (mu = 2.80, M = 3, sd = 1.05) and Q3 (mu = 4.35, M = 5, sd = 0.93) show that the difficulty for the test was appropriate, but most subject tended to give the answer that it was a little too difficult. The answers of Q2 could have also been a result of the formulation of the question. The possible range of answers only included that the lectures were not too difficult and that they were too difficult. If the question would have asked for the difficulty with 1 meaning that it was too easy and 5 meaning that it was too difficult different answering patterns might have emerged.[24]

In the following part, we will focus on the differences in answers for the two groups. It should be noted here that question 10, Q10, will be excluded. Several participants reported either not knowing the word "sentient" or being unsure about its meaning. The question was supposed to measure the anthropomorphism felt by the participants. Several participants misinterpreted the term to be synonymous with "sentimental". Therefore, it seemed reasonable to exclude this question, since we cannot be sure how the other participants, who did not mention it, understood the question.

In question 4, Q4, there is a significant difference in the answers between the two groups (t = 2.2094, df = 14.302, p-value = 0.04393). The negative group reported that it affected them emotionally (mu = 3.9 , M = 4, sd = 0.57), while the positive group showed moderate emotional reactions (mu = 3.1, M = 3.5, sd = 0.99). These findings are significantly different from 0, implying that the robotic feedback caused emotional reactions in both groups.

Question 4.5, Q4.5, was a text field, which instructed the participants to describe their emotional reactions. 14 Participants (8 in the positive group, 6 in the negative group) gave an explanation. In the negative group, the most common reaction was to feel uncomfortable and attacked. While this was a common thread it should be noted, that it is difficult to summarize all these explanations since they varied. For example, one participant wrote "When it was angry because I wanted to repeat a lecture, I felt stupid and wanted to do better." This participant took the criticism to heart and tried to appease the robot by trying harder. Another participant wrote "I was a little trotzig, felt treated unnecessarily harsh, had the impuls [sic] to just skip repeating to be done more quickly." This participant admitted in the feedback that they did not follow the impulse to skip repeating. Instead, they did the opposite and took retake on purpose, only because it seemed to defy the robot's expectations. This perception of the robot not wanting the subject to retake a lesson was also described by other subjects.

In the positive group, the feedback was mostly perceived as positive. This is also shown in Q8 ("8/10 Did you find the robot to be sympathetic?") in which there was a significant difference between the groups (t = -3.7264, df = 14.014, p-value = 0.002253). The positive group perceived the robot to be much more sympathetic (mu = 3.7) than the negative group (mu = 1.7). Only a few participants in the positive group found the robotic reactions to be annoying. It was mentioned that the repetitiveness was making the performance unrealistic.

Q6 ("6/10 Were the reactions of the robot helpful and or beneficial for your learning experience?") and Q7 ("7/10 Overall, did you find the robot to be a good teacher?") did not show significant differences between the groups, for Q6 (t = -1.1294, df = 17.646, p-value = 0.2738) and for Q7 (t = -1.8, df = 16.691, p-value = 0.08996). This is notable insofar as there was an extreme difference in answering behavior in Q8 but not in Q6 and Q7. This might show that the participant did not believe that the negative reactions were bad, although unpleasant. There it seems to suggest, that a good teacher does not need to be liked. In fact, the strongest factor in explaining Q7 is Q6, since they correlate significantly (t = 2.7456, df = 18, p-value = 0.0133). This means people who found the reactions to be beneficial also tended to judge the robot to be a good teacher. There is a correlation between Q7 and Q9 ("9/10 Do you feel like the robot cared about your success?") of 0.4098013 but it fell short of being significant (t = 1.906, df = 18, p-value = 0.07274).

We did not find any correlations between emotional engagement and performance. In fact, participants who reported similar reactions did not score equal. This might be due to two reasons: firstly the individuality and secondly the steadiness of performance. We will elaborate on these two hypothesized factors.

What we mean by individuality is, that individual traits strongly influence every part of the experiment. Not just the performance in the quiz but also how the robotic feedback is perceived and how it influences the subject. It is not hard to believe that some subjects are just better at memorizing lectures which they hear. Our findings suggest on top of that, that reactions to robotic feedback are just as individual. Some feel wrongfully attacked by the robot, others feel like it is their fault. Some of those who feel attacked, feel motivated to prove the robot wrong and show that they can do it, others withdraw and lose motivation. This is surely one of the most interesting findings. We hypothesis that personality is an important factor. It would be really a fruitful enterprise to add a personality test to the experiment and to see whether certain reactions can be correlated with personality traits.

The second factor we hypothesize is what we call steadiness of performance. This means that we hypothesize that the strongest influence on performance in the test is the mental

[24] The R-Code was uploaded to the GitHub Repository: https://github.com/roboprofs/documentation/blob/main/analysis/qualitative_analysis.Rmd

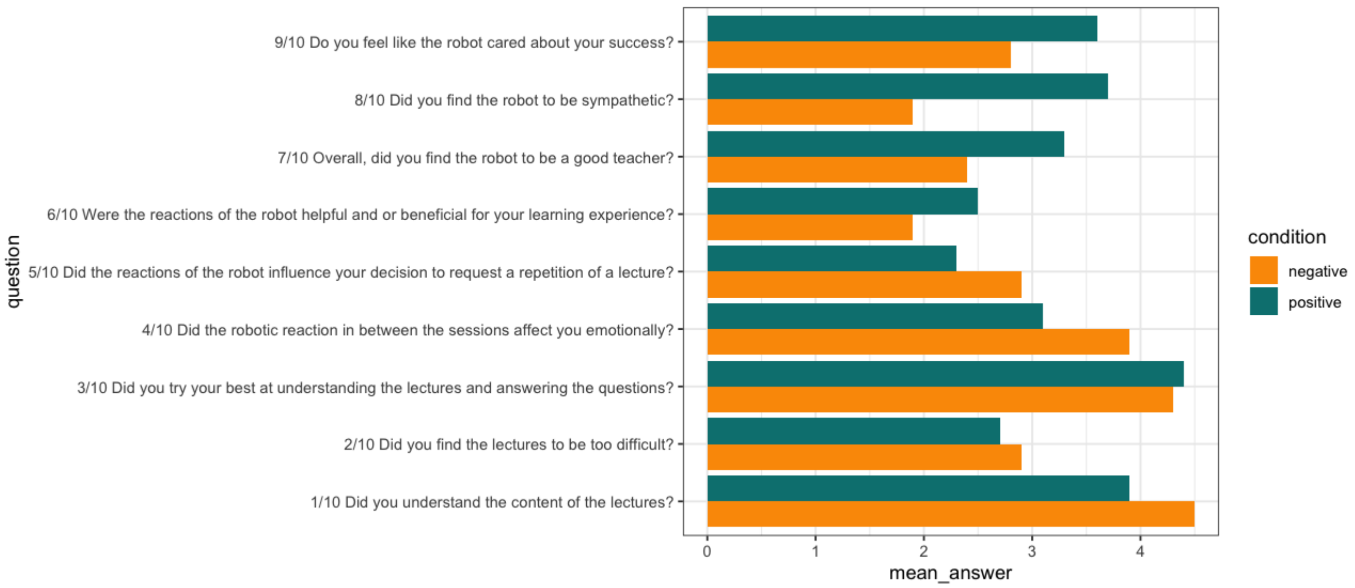


Fig. 11. Mean answers to the questionnaire for both groups

ability of the participant and therefore outside of our reach. All the factors we were able to manipulate will only slightly affect the performance of the subject. This is also supported by a power analysis we did, which predicted, that in order to achieve a statistical power of 80 percent, we would have needed 205 participants.[25]

This means that the overlap in performance between the groups is big and the difference caused by the different conditions is not extremely big. Therefore a lot of subjects would be needed to be able to find significant proof for a difference between the groups.

205 participants are quite a larger number. Therefore, it might be better to add a general memory test before the experiment to be able to use this data as a reference to accurately measure the influence of the feedback.

V. Discussion

A. Conclusion

The technical execution and schedule on the experiment days went as planned and we were able to motivate 20 participants to take part in our experiment. We were able to conclude from the verbal feedback that our research on social cues was paying off, as they elicited the intended responses from the participants. For example, individuals in the negative condition felt hassled by the robot (proximity). It was interesting to see how people reacted to the robot in different situations/ conditions. The synthetic voice was also worthwhile, as it was perceived as very authentic and emotional. The performance hypothesis was not supported by our experiment data, since the participants of the negative condition were slightly better in answering questions regarding the lessons. The repetition hypothesis was supported by our experiment data since the participants of the positive condition repeated the lessons more often than the participants in the negative condition.

B. Shortcomings

During the semester and the experiment, a few problems occurred that were problems we had no impact on. For example, none of us had prior knowledge of Unity and Blender and therefore we underestimated the amount of work needed for the animations. We all needed to get a basic understanding of the necessary tools and how to use them. A problem during the experiment was that the FFP2 masks that the participants had to wear were perceived as distracting and caused some of them to skip quickly. It would be interesting if the FFP2 masks had a measurable impact on the number of repetitions. We only know what the people told us and what we noticed during the experiments. Also, the subjects' English skills varied (e.g., many did not know the word sentient). We think that all people understood most of the content correctly anyways. In addition to that, we had the feeling that some of the people were simply good at memorizing things so that they were better in the content-related questionnaire no matter in which condition they were. We know that this would not be a problem if we would have enough (representative) participants. Maybe we could have done a standardized memory test beforehand as a reference, but this seemed/ seems to be really sophisticated for our demands. Other problems with so few participants were that the lesson order was not contributed evenly (See Figure 12).

[25] t tests - Means: Difference between two independent means (two groups) Analysis: A priori: Compute required sample size Input: Tail(s) = Two Effect size d = 0,3897376 alpha err prob = 0,05 Power (1-beta err prob) = 0,8 Allocation ratio N2/N1 = 1 Output: Noncentrality parameter = 2,8239172 Critical t = 1,9714347 Df = 208 Sample size group 1 = 105 Sample size group 2 = 105 Total sample size = 210 Actual power = 0,8025862

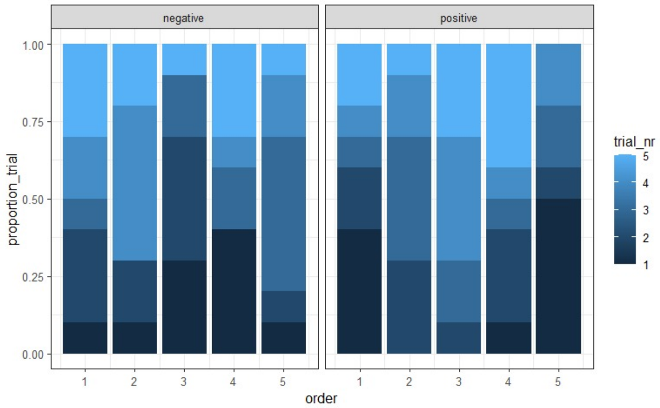


Fig. 12. Distribution of the lessons in each trial

However, there were of course a few mistakes in the experiment itself that we hadn't noticed before: For example, the robot asked,

> *Did you understand everything or should I repeat the lecture?*

which was ambiguous and the subjects did not know which question to answer. This resulted in several unintended repetitions. After noticing the ambiguity, we pointed to this situation to each participant before the experiment started. These extra instructions solved the problem, but the next time we would pay more attention to situations that could be perceived as ambiguous. Another thing we noticed was that the Citizenship test was probably slightly too difficult and people often just made educated guesses. The best participant had 80%, the average had about 41%. That is clearly more than the 25% that a random selection would likely result in. But since the questions were differently difficult some seemed to be randomly selected. The next time we would reduce the number of new names and instead focus more on number facts and relation facts. Another problem was that our open questions were not answered in enough detail. Next time we would focus more on qualitative questions in the questionnaire and write a clearer description, of what we are aiming for. For example by adding a desired number of sentences:

> *4/10 Did the robotic reaction in between the sessions affect you emotionally? 4.5/10 If yes in what way? (Please answer in at least 3 full sentences.)*

Due to the time pressure of the course, we decided to make no test run with unbiased persons. Looking back, we would prioritize a test run more because we think most of the problems would be clear after a test run. (For example, the complexity of the questionnaire, ambiguity of the question) Because we would need more time to do a test run we would choose a less theoretical and more practical approach to the project next time: We tried to make everything as scientific as possible, searched papers for nearly everything, but in the end, we had time pressure. The time could be better spent on further work of the animations, the robot feedback, or as mentioned a test run. The feedback of the robot did not change within the conditions. So if a participant repeated more than one lecture the feedback was the same all the time. If we would have had more time, we could have let the feedback increase in its intensity (for example like this: *first repetition - neutral, second repetition - angry, third repetition - very angry, fourth repetition - offensive, fifth repetition - very offensive*) This would help to make the scenario more realistic.

C. Outlook

There are several ways and areas one could expand our experiment. (If one wants to make our experiment only more professional, see Section [V-B]) For example, the experiment design could focus more on the interaction between the human and the robot. An idea to do that is to add a third part to the experiment, which is a personality test. That would help to get a closer understanding of how the interaction with the robot changes the feelings and understanding of the participants. It could be interesting to see how different personality types react to the different condition feedback. We think the results of the hypotheses could correlate with certain personality types. This idea came to us, because we noticed that the same feedback of the robot can evoke different feelings in the participants.

REFERENCES

[1] Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007. Special issue on Information technology.

[2] Bipika Amatya. Emotional speech from machine. 2020.

[3] Lisa Feldman Barrett. *How emotions are made*. Pan Books, London, England, 2018.

[4] Casey Bennett and S. Sabanovic. Deriving minimal features for human-like facial expressions in robotic faces. *International Journal of Social Robotics*, 6:367–381, 08 2014.

[5] Cindy Bethel and Robin Murphy. Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38:83 – 92, 02 2008.

[6] LaVonda Brown and Ayanna M Howard. Engaging children in math education using a socially interactive humanoid robot. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 183–188. IEEE, 2013.

[7] Felix Burkhardt and Nick Campbell. Emotional speech synthesis 20. *The oxford handbook of affective computing*, page 286, 2014.

[8] Albert Causo, Giang Vo, I-Ming Chen, and Song Yeo. Design of robots used as education companion and tutor. *Mechanisms and Machine Science*, 37:75–84, 09 2015.

[9] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28:117–139, 01 2004.

[10] Stephanie Embgen, Matthias Luber, Christian Becker-Asano, Marco Ragni, Vanessa Evers, and Kai Arras. Robot-specific social cues in emotional body language. pages 1019–1025, 09 2012.

[11] Stephen Fiore, Travis Wiltshire, Emilio Lobato, Florian Jentsch, Wesley Huang, and Benjamin Axelrod. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, 4:859, 2013.

[12] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[13] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. Characterizing the design space of rendered robot faces. pages 96–104, 02 2018.

[14] Petri Laukka, Patrik Juslin, and Roberto Bresin. A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5):633–653, 08 2005.

[15] Paul J. Leach, Rich Salz, and Michael H. Mealling. A Universally Unique IDentifier (UUID) URN Namespace. RFC 4122, July 2005.

[16] Hagen Lehmann, Alessandro Roncone, Ugo Pattacini, and Giorgio Metta. Physiologically inspired blinking behavior for a humanoid robot. volume 9979, pages 83–93, 11 2016.

[17] Max Louwerse, Arthur Graesser, Shulan Lu, and Heather Mitchell. Social cues in animated conversational agents. *Applied Cognitive Psychology*, 19:693 – 704, 09 2005.

[18] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[19] Loh Ngiik Hoon. Development of real-time lip sync animation framework based on viseme human speech. *Archives of Design Research*, 112:19, 11 2014.

[20] Antonio Rui Ferreira Rebordao, Mostafa Al Masum Shaikh, Keikichi Hirose, and Nobuaki Minematsu. How to improve tts systems for emotional expressivity. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[21] Alwin Rooij, Joost Broekens, and Maarten Lamers. Abstract expressions of affect. *International Journal of Synthetic Emotions*, 4:1–32, 01 2013.

[22] Mostafa Shaikh, Antonio Rebordao, Keikichi Hirose, and Mitsuru Ishizuka. Emotional speech synthesis by sensing affective information from text. 09 2009.

[23] Noé Tits. A methodology for controlling the emotional expressiveness in synthetic speech - a deep learning approach. pages 1–5, 09 2019.

[24] Johan Verwey and Edwin Blake. The influence of lip animation on the perception of speech in virtual environments. 01 2005.

[25] Joseph Weizenbaum. ELIZA — a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 26(1):23–28, jan 1983.

[26] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.