Social Engagement in Children-Robot Interaction Over Multiple Sessions

by

Sofia Kostakonti

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday April 8, 2024 at 10:00 AM.

Student number:	5494826	
Project duration:	April 18, 2023 – April 8, 2024	
Thesis committee:	Dr. C. R. M. M. Oertel,	TU Delft, supervisor
	Prof. Dr. M. A. Neerincx,	TU Delft, supervisor

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

With the advancement of technology and the integration of Large Language Models, interactive robots are entrusted with more significant tasks that contribute to human well-being. This transition is particularly significant in educational settings, where social robots assume diverse roles such as teacher, tutor, and instructional tool, aiming to enhance learning experiences. Engagement emerges as a central theme in Human-Robot Interaction studies, especially within pedagogical contexts, where it has a significant impact on learning outcomes. However, optimizing engagement rather than maximizing it may yield superior learning results. This thesis investigates the correlation between the employment of engagement strategies and learning outcomes in prompt-driven discussions with robots, focusing on facilitating self-reflection and understanding personal values among children. Due to limitations with the target population, a total of 55 university students conversed with the robot in two different sessions, where they discussed different situations, how they would react in each, and the reasoning behind their behavior, using the Schwartz values as a basis for the choices. The participants were divided into two conditions, to examine the effect of techniques such as motivational interviewing, cues to images, feedback, and the use of a memory model on the quality of their arguments and their ability to recollect the interaction. Results suggest that the strategies lead to significantly more reflective arguments for the first session, but they equalize in the second interaction. No significant effect was found between condition and participants' identifying their value profile which led to an examination of the assumptions made by the robot. Participants who agreed with the assumptions were more likely to identify their values correctly, but, generally, the value model was unable to fully capture the intricacies of human motivation. Participants found the robot equally likable, a possible result of the novelty effect. On recollection quality, the data shows that the effect of the session is too strong to allow room for the condition. Overall, this study contributes valuable insights into the influence of engagement strategies on learning outcomes in educational interactions, offering guidance for designing more effective and engaging interactions in the future.

Preface

This thesis marks the completion of my degree and the end of my journey at TU Delft. When I first joined the program, I was unsure of what my particular area of interest would be, and I planned to seize as many opportunities as I could to find the one I was passionate about. Being a part of the initial study that this one is based on, highlighted for me my interest in humanity above technology and how enhancing human-robot interactions is imperative for the future of a human-centered society. The process of designing and implementing the experiment, although extremely rewarding, was not without its challenges. Trying to convince schools to trust me and collaborate with me taught me perseverance and patience but the community that rallied and helped me complete my experiment reminded me why robots will never replace humans.

At this point, I'd like to thank all the people who supported me during the past year and long before that. My supervisors, Catharine Oertel, and Mark Neerincx, for their support and guidance, and especially Catha for all the encouraging words and hugs when everything seemed overwhelmingly impossible. I would also like to thank Franzesca Burger for letting me join her study, learn from her, and for providing me with all her insight, advice, and comradery on running an experience on a large scale. Many thanks to the rest of the members of the Interactive Intelligence research group with whom I talked and brainstormed ideas. I'd like to thank Jean-Paul Smit, for helping with finding participants, keeping me company, and reviewing the argument labels. Finally, I'd like to thank everyone who participated in this research experiment.

Outside of my academic environment, I had so much support from friends and family that I would truly not have made it without them. A very special thanks to my parents who have supported me throughout all my life in everything I wanted to do and have helped get me to where I am.

Σας ευχαριστώ.

Sofia Kostakonti
Delft, April 2024

Contents

Chapter 1

Introduction

As technology advances in the realm of interactive robots, the objective of research is shifting from simply facilitating conversations, to ensuring their value and benefit to humans. The introduction of Large Language Models (LLM) has enhanced natural language processing, enabling robots to understand complex human speech and respond more effectively. This transition becomes even more crucial when we take into account how these advancements affect children within educational settings. In recent years, social robots have been utilized in "in the wild" studies, meaning deployed in actual classrooms and not in a controlled environment, and have assumed many different roles, such as teacher, teaching assistant, tutor, instructional tool, and peer learner[1]. In terms of goals, robots have been employed to facilitate tutoring[2], language learning[3], as well as interventions in special education[4], to name a few.

A common theme in Human-Robot Interaction studies, particularly in the pedagogical context, is engagement during the interaction. Engagement has many definitions, but a widely accepted one, that is also adopted for this study, is the one by Snider[5]: "Engagement is the process by which the individuals involved in an interaction start, maintain and end their perceived connection to each other". A further distinction is often made between task engagement and social engagement [6], which refers to the receiver of the user's attention. The user being engaged with the agent is referred to as social engagement, while being engaged with the agent, as well as the task, is referred to as task engagement. In studies where

an activity is part of the interaction, task engagement is more commonly studied, whereas social engagement is more prominent in studies where the activity is the conversation, and often includes an affective component. The importance of engagement is highlighted when considering that students are more motivated and engaged and achieve better results in terms of learning outcomes when they are instructed through active learning, meaning involving students in the learning process[7]. The contexts within which robots have been employed in classrooms are inherently interactive and directly involve the students actively participating in the interaction. Therefore, it can be inferred that engagement is directly correlated with the accomplishment of the learning outcomes.

Multiple measures have been used to assess and quantify engagement, including self-report, manually annotated media such as video and voice recordings, multi-modal machine-learning models, and even more invasive methods such as wearable devices and monitoring brain activity through EEG[8]. Whereas most studies consider maximum engagement the best for the interactions, Nasir et al. [9] challenge the link between maximizing engagement, in terms of human-annotated media, and increased learning outcomes, and they instead propose that optimizing engagement should be the goal. They define the term 'Productive Engagement' as "the level of engagement that maximizes learning" and explore how successful learning can be predicted by machine-learning models. Following that line of thought, this study will attempt to provoke and optimize engagement, not measured by any of the aforementioned methods but by assessing the learning outcomes.

Before defining the learning outcomes, it is important to establish the context of this research. The experiment will involve a two-session prompt-driven discussion with a robot about personal values and decision-making in different scenarios. A similar experiment first took place in the context of a pilot study earlier in the year and several things were adapted from that design. The study was performed by Franziska Burger, who designed the experiment around personal values, as expressed by Schwartz [10]. These values represent basic human motivations and understanding one's values can increase confidence in decision-making, and enhance self-awareness and relationships with others. The goal was to study how a memory model for a conversational agent could assist children in navigating various

situations within the school context and engaging in self-reflection to determine their values. The children were presented with a binary behavior choice, where each option represented a personal value. The premise of the interaction shall remain the same and more information regarding the design and the implementation is presented in following sections.

Particularly in interventions where learning is the objective, engagement is often studied concurrently with the learning outcomes of the interaction, and information retention is frequently considered a learning outcome and an indicator of engaged participants[11, 12]. It is an intuitive assumption that the more engaged people are in a conversation or a task, the more they will be able to recall it after the fact. Subsequently, information recall about the scenarios discussed will be considered one of the learning outcomes of this interaction.

However, another learning goal of the experiment is to promote self-reflection and understanding of one's motivation behind decisions. During the discussion, participants are prompted to argue about their choices. The quality of those arguments, or rather their improvement, will be the second learning outcome that this study will try to maximize.

This research, along with the pilot study that preceded it, is part of the ePartners4all Project. The objective of the project is to address health promotion challenges in children, particularly those facing economic disadvantages or with specific health needs. As part of this innovative initiative, this study focuses on creating an interaction that revolves around personal values and the understanding of inner motivation. It is an opportunity to enhance self-awareness and alignment with one's goals, which are key for motivation, personal growth, and good mental health. This utilization of tailored and interactive e-health solutions can eliminate the need for one-on-one time with experts who are severely understaffed in classrooms, and give access to children who would otherwise be deprived of it.

Chapter 2

Background

2.1 Schwartz values

Introducing the Schwartz Values is crucial for setting the frame of the experiment as it serves as the foundation of the interaction for evaluating participant engagement and understanding. The Schwartz Values are a set of universal values, developed by psychologist Shalom H. Schwartz [10], which represent basic human motivations. Schwartz proposed a theory of these motivations grouped into a model of 10 core values organized into a circular structure known as the "circumplex model" (Figure 2.1). These values are thought to be cross-culturally relevant and have been applied in numerous research projects to understand the motivations and priorities of individuals in diverse societies.

The Schwartz model includes the following ten fundamental values, along with their main goals:

1. Self-Direction: emphasizing independent thought and action, creativity, and exploration of one's own ideas
2. Stimulation: seeking excitement, novelty, and challenge in life
3. Hedonism: pursuing pleasure and sensuous gratification for oneself
4. Achievement: striving for personal success through demonstrating competence according to social standards

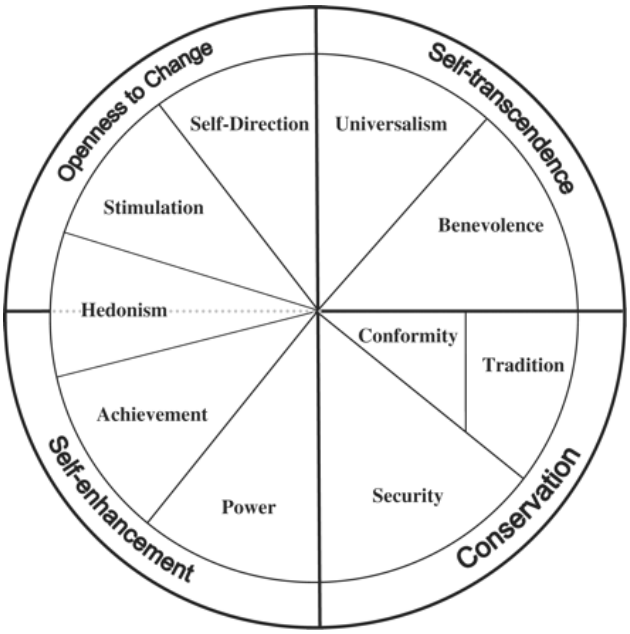

Figure 2.1: Schwartz circumplex model[13]

5. Power: seeking influence or control over people and resources
6. Security: desiring safety, stability, and harmony of society, of relationships, and of self
7. Conformity: restraining actions and impulses that are likely to upset others and violate social norms
8. Tradition: respecting and adhering to cultural or religious customs and ideas
9. Benevolence: valuing and protecting the well-being of those with whom one is in close personal contact (the 'in-group')
10. Universalism: emphasizing understanding, appreciation, tolerance, and protection for the welfare of all people and nature

The circumplex model arranges the values in a circular structure, where more compatible values appear closer together in the circle, whereas those opposite each other are in tension. Individuals may prioritize and internalize these values differently, and the importance given to each value depends on the individual and their cultural background.

2.2 Self-Determination Theory

When attempting to optimize engagement, it is important to consider what is engagement and where it stems from. During an interaction, engagement manifests as observable behavior, while motivation is the underlying reason for such behavior [14]. Therefore, to enhance engagement, it is first necessary to reinforce the motivational factors that prompt individuals to actively participate in the interaction. One theoretical framework that unravels these factors in distinct categories is the Self-Determination theory. Self-Determination Theory (SDT) is a theory formulated by Deci and Ryan[15], which aims at explaining human motivation and behavior. According to SDT, there are three basic psychological needs that drive people: competence (having the skills to perform a certain task), relatedness (feeling a sense of closeness and support by those around them), and autonomy (feeling independent and in control of their behavior). When individuals have self-determined reasons for their actions, they are more likely to be engaged in those actions. In the following sections, an exploration of the approaches employed to address the three basic needs and achieve higher motivation is undertaken.

2.2.1 Autonomy

Motivational Interviewing

Motivational Interviewing (MI) is a counseling approach for supporting behavior change formulated by Miller and Rollnick[16]. Although this method aims to resolve ambivalence and prepare people for change, in the context of this thesis, MI techniques will be employed to help them gain a deeper understanding of their own motivations and thoughts, which will hopefully promote self-awareness and better argumentation about their choices. MI has been applied in multiple studies where an agent was interacting with the participants, mainly to provoke behavior changes for health-related advantages. Schulman et al.[17] designed a conversational agent to promote long-term health behavior change by implementing MI techniques, producing results that were rated significantly high by both experts and users. Kanaoka and Mutlu [18] used MI and a physically embodied agent to increase physical activ-

ity. While the study did not provide the expected results, they attribute it to inconsistencies in the dialogue flow and speech recognition that were necessary for an adaptive interaction. More recently, Samrose and Hoque[19] developed a motivational interviewing chatbot to improve conversational skills and the user's confidence. The results of their study show that participants followed the MI agent's suggestions and feedback more often than those who interacted with the non-MI agent. Consequently, it can be surmised that the application of MI techniques through an agent can yield positive effects on users. Within an educational context, Standage et al.[20], albeit without a robotic agent, showed that providing an autonomy-supportive environment can have significant effects on self-determined motivation. This study aims to integrate these findings and explore the feasibility of applying MI techniques in decision-making scenarios to improve argumentation quality.

2.2.2 Relatedness

Social Penetration Theory

The Social Penetration Theory (*SPT*), developed by Altman and Taylor [21], describes the process of information exchange in interpersonal relationships and how relationships progress to more intimate and meaningful connections through self-disclosure[22]. The more a relationship develops, the more it grows in depth, which, according to Altman and Taylor, allows for the sharing of values and beliefs. In human-robot interaction, self-disclosure can enhance the human-like qualities of the robot, making it more relatable, strengthening the bond between participant and agent, and providing emotional support and encouragement. As a result, the participants can internalize the motivations of the agent, and therefore their own self-motivation will be enhanced.

Burger et al.[23] studied the effect of self-disclosure on relatedness, utilizing an application for diabetic children where an avatar accompanied them, prompting self-disclosure questions. The results indicated that the ratio of active disclosures, meaning cases where the participants responded to the robot in self-disclosure attempts, to total disclosures, significantly predicted the feeling of relatedness. Furthermore, they found that relatedness was also significantly linked with consistency and average time of activity, which can be evidence

of their intrinsic motivation to use the application. Another study that performed a second language learning experiment[24] showed that attempting to evoke relatedness during the interaction is more effective than when excluding it.

2.2.3 Competence

Feedback

To enhance the perceived feeling of competence during the interaction, an important factor is the feedback provided, which can be in the form of motivational feedback, positive feedback, and gestures. Motivational feedback, in the context of motivational interviewing or feedback to the speaker's response, aims at re-engaging the participants if that is deemed necessary. The goal is to reignite their interest and active participation in the conversation.

Positive feedback is not about rewarding correct responses, since there are no wrong or right answers in the design for this study. Instead, giving non-generic praise to the participants based on their responses will be imperative, specifically if they demonstrate depth and thoughtful consideration, in order to encourage similar argumentation. Research has shown that non-generic praise can have a more significant impact on intrinsic motivation than generic praise[25, 26]. Additionally, positive feedback can be given in the next interactions, based on the participants' overall involvement in the previous sessions, further promoting their motivation and engagement.

Lastly, gestures can be utilized in two forms during the interactions. Firstly, as positive feedback, such as a thumbs-up gesture[27], to provide affirmation and encouragement. Secondly, gestures as deictic means, directing the participants' attention towards elements that require their focus.

2.3 Memory

A significant advantage of long-term interaction is the ability to incorporate memory into the robot and have it recall information established in previous sessions. Personalizing the interaction has been shown to maintain the willingness to continue with the interaction across

multiple sessions[28]. Some ways this can be achieved is through mentioning the participant's name[29] or recalling shared experiences[30], where moments from past interactions are mentioned in conversation to increase affect and deepen the relationship. Shared experiences are also relevant in the context of motivational interviewing since they can be used to adjust the feedback in a more personalized and meaningful manner[31]. Since the personalized feedback takes into account shared memories, it can be more relevant and engaging to the participants.

Dialogue templates are a common way to personalize the interaction where blanks are filled with the information provided by the participants [28, 32]. Kruijff-Korbayova et al.[33] apply a persistent user model to signal familiarity with the user, mentioning name and references to previous interactions and performance in the activity. Fu et al.[**Fuetal2021**] share experiences to evoke empathy. Those experiences are injected into the dialogue flow and reference pre-determined memories of the robot that do not include the subject, however, they are unrelated to the participants. This is mitigated in a similar study[35] where the dialogue input is analyzed and stored in a database. The robot then proceeds to match the memory with previous experiences and generates utterances based on them. Leite, Pereira, and Lehman[36] use persistent memory to establish familiarity and rapport. The robot has access to information provided by the participants out of view and that it can use in templated dialogue to relate more to the subjects. In the study performed by Churamani et al. [37], a model of the robot's world is created. The interactions center around the participants teaching the robot the whereabouts of specific objects, and a personalization condition is defined for which a language module is developed. This module utilizes Natural Language Understanding (NLU) and Named Entity Recognition (NER) to extract information from the user's utterances, then updates the knowledge base, and generates relative dialogue with Natural Language Generation (NLG) and template sentences.

A different way to incorporate the memory and shared experiences is through adapting the interaction based on different situations. Ahmad, Mubin, and Orlando[38] designed a game of snake and ladders between children and a robot, where the robot stores various relevant information about the children, their friends, their game performance, and their

reactions. In a series of related papers[**Ahmadetal2022**, 39], the robot's response unit is refined to include modules for emotional event calculation, memory mechanism generation, and a behavior selection unit. Events are stored in relation to the emotions evoked at the time and an appropriate response is chosen.

A different approach is a simulation of a human long-term memory system. In cognitive science theory, the memory system is comprised of both episodic and semantic memory[40]. Episodic memory refers to recollections about specific events, and the tempo-spatial connections between them, and is mostly related to personal experiences. Semantic memory, on the other hand, deals with general knowledge that a person possesses about the world, without it being related to a particular event[41]. Ho et al.[42] propose the modeling of short-term and long-term memory. Short-term memory includes the active, completed, or violated goals of the interaction, while long-term memory is comprised of general event representations, world knowledge, and autobiographic memory, which are constantly examined against the short-term goals of the interaction. Yumak and Thalmann[43] also simulate episodic long-term and short-term memory. As part of the long-term interaction, a face recognition module is incorporated and a Hierarchical Task Network (HTN) planner is used for memory retrieval. Experiences are first stored in the short-term memory, and then only those that are emotionally salient are chosen to be stored in the long-term memory to be mentioned at a later time. A machine with a human-like memory system that includes both episodic and semantic memory has been implemented before[44] and has shown to outperform similar machines without this structure in their system. For this study, a similar system is implemented, combining information from both memories to achieve a more realistic information callback in the second interaction.

2.4 Research Question

The main objective of this study is to research the impact of the discussion with the robot on the understanding of the chosen topic and to what degree engagement strategies and memory can support this goal. To explore this objective, the following hypotheses are drawn:

H1: Participants in the memory condition will have a better quality of argumentation,

especially in the second interaction. The strategies employed in the memory condition are expected to enhance the participants' cognitive processes and assist the participants in providing better quality of argumentation. Introducing the memory aspect during the second interaction will further improve the arguments' quality.

H2: Participants in the memory condition are more likely to agree with the robot's value profile about them.

By the end of the experiment, participants in the memory condition are expected to have a better understanding of their personal values. That would lead them to identify more easily the value profile attributed to them by the robot, and with greater confidence.

H3: Participants whose arguments reflect the assumed value of their choice are more likely to identify the robot's value profile correctly.

Regardless of the condition, participants whose arguments better match the assumptions of the robot about their behavior will be able to correctly identify the value profile attributed to them by the robot.

H4: Participants in the memory condition will find the robot more likable and lifelike.

In the memory condition, there are strategies to increase the relatedness of the robot, which will enhance the overall interaction quality, leading to a more positive perception of the robot's likability and lifelikeness among participants.

H5: Participants in the memory condition will have better recollection quality than those in the baseline condition.

Participants in the memory condition are expected to be more engaged during the interaction, leading to improved retention of information and more accurate recounting of moments from the conversation.

Chapter 3

Robot Design

3.1 Discussion Topic

The topic of the conversation between the robot and participants is several school-related scenarios where the subjects are presented with a binary behavior choice and each option reflects a personal value. The personal values can be one of the following: Benevolence, Achievement, Conformity, and Self-Direction, and they each represent one of the four quadrants from the Schwartz circumplex model (Section 2.1). Thus each scenario has a value conflict (e.g. Benevolence vs. Self-Direction) and all possible combinations(6) are presented within the first session.

However, most people don't necessarily react the same way in every situation, and they might choose one value over the other depending on the context. Therefore, an additional variable is introduced, the situational characteristics which are derived from the DIAMONDS taxonomy [45], and used to make the agent aware of the context in each scenario. For this experiment, the 3 most psychologically relevant characteristics were chosen to be included, namely Duty, Intellect, and Adversity, and most of the combinations of their presence or absence are taken into account (6/8 combinations).

A few examples of the scenarios discussed, along with their value conflict, and situational characteristics are presented in Table 3.1. A full list can be found in Table A.1 of Appendix A.

#	**Value Conflict**	**D.I.A.**	**Scenario description**
1	Self-Direction vs Benevolence	0,0,1	you play tag and your friend denies that you tapped her
2	Achievement vs Benevolence	0,1,0	the person sitting next to you can't concentrate well during a math course and starts distracting you
3	Conformity vs Self-Direction	0,1,1	you have to prepare a group presentation but you disagree with the topic chosen by the group
4	Conformity vs Achievement	1,1,1	your friends would like to go to an environment strike but the professor also wanted to teach an important lesson

Table 3.1: Examples of scenarios to be discussed. Situational characteristics (Duty, Adversity, Intellect) are denoted with 1 if present and 0 if absent.

3.2 Cue to Images

In the pilot study, images were displayed on a screen in front of the participants to depict the scenarios and give a better understanding of the situation. The images would include two pictures, each representing one of the behavior choices, and the main character was introduced at the beginning and remained consistent throughout all the situations. The background and the amount of other people in the pictures varied depending on the scenario. Some examples of those images can be found in Figures 3.1 and 3.2.

Figure 3.1: "A friend accuses you of stealing her pen. Do you defend yourself or help find it?"

Figure 3.2: "Your friends want to go to a protest but the professor wants to teach an important lesson"

Initially, this was a medium to assist participants in following the scenario, especially since the pilot study was done on children. However, research stemming from this experiment showed that participants who could reproduce 'high quality' recollections of a scenario, spent more time looking at the screen during the specific scenarios[46]. Therefore, a research

opportunity arises to try and reverse the finding. The assumption would be that if participants spend more time looking at the screen, they will remember the scenarios better after the interaction. To challenge this assumption, the robot cues towards the images at the beginning of each scenario using phrases such as *"Look at the screen! It will help you understand the scenario better."* or *"You can now study the pictures in front of you showing the different options."* with the expectation that participants will indeed pay more attention to the pictures.

3.3 Memory System

For this experiment, the existence of a memory system that utilizes both episodic and semantic memory to store and retrieve relevant information during the interaction is imperative, as explained in section 2.3.

The episodic memory stores data related to the scenarios, the decisions each participant made, and the reasoning behind them. These are later mentioned in the conversation to remind the participants of the specific event and ask them to further argue about their choices. This is expected to elicit more positive reactions about the robot and lead to participants perceiving it as better liked and accepted[34]. Additionally, the participants' names are stored in episodic memory, utilized for the same reason at the beginning and end of the interactions. Similar studies [33, 38, 37] have used first names, in addition to shared memories, to implement a personalized interaction with mostly great results regarding the robot's likeability and intelligence. The method of storing these values is a simple .csv file with records for each of the scenarios discussed.

The semantic memory allows for the robot to make deductions about the value profile of each participant based on their overall interaction and their decision-making process throughout all of the scenarios. For the storing of this abstract information, the participants' choices and the words used to motivate them are used to build a Bayesian network that links the situational variables to the specific values. A representation of the network can be seen in Figure 3.3.

At the start of session 1, the network is initialized, and probabilities are assigned initial

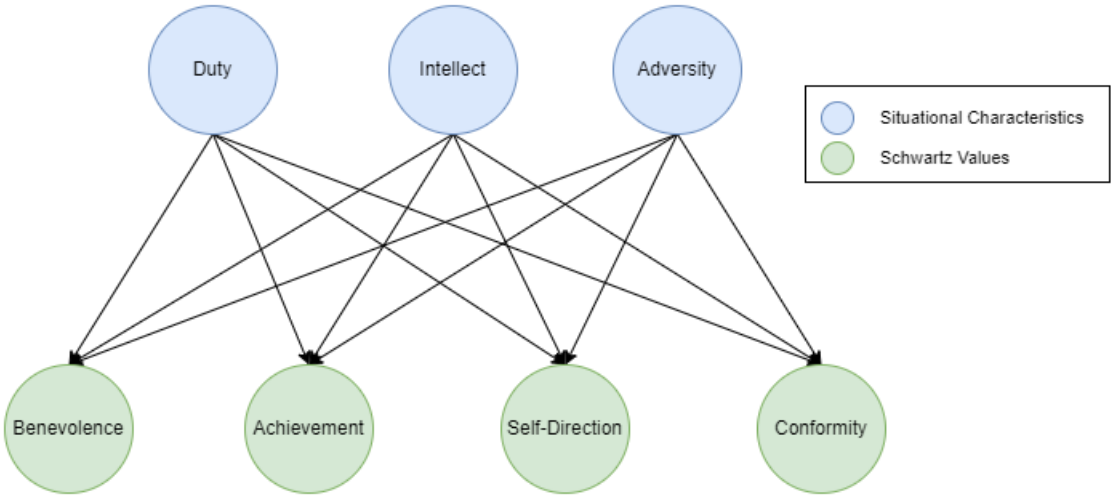

Figure 3.3: Graph representation of the Bayesian network for the semantic memory

values. These values are informed by the situational characteristics associated with each Schwartz value and the participant's personal ranking of the Schwartz values at the beginning of session 1. More specifically, each Schwartz value is intuitively linked with the situational characteristics. For example, it can be deduced that "Achievement" is positively linked to "Intellect" but usually negatively linked to "Duty", while "Adversity" has no significant effect. On the other hand, "Benevolence" is the exact opposite. Similarly, "Conformity" is positively linked to "Duty" and negatively linked to "Adversity", with no obvious links to "Intellect", while, for "Self-Direction", it can be said the opposite: "Duty" can affect it negatively, and "Adversity" in a positive way. Since these are intuitive rules, the change of the priors is relatively small, where they are set to 0.4 (or 0.6) in the case of negatively (or positively) affected by the two linked characteristics, and they are set to 0.5 if they counteract each other.

Since the scenarios and the time to update the network are limited, a piece of stronger information was considered as a prior for the initialization. Before the first interaction, participants are asked to rank several Schwartz values, depending on how important they find them. According to their position in the ranking, and using the value provided by

the situational characteristics computations, the priors are changed once more to better reflect the participant's view of each value. If the value was ranked at positions 1-3, a small number, appropriate to the ranking, was added to the prior, whereas at positions 4-6, it was subtracted.

The Bayesian network updates every time there is new information about the participant, meaning when a participant's decision and its argumentation are provided. The data that the network uses as input is the context of the scenario (the combination of situational characteristics), the subject's chosen and discarded values, and their argument. First, the argument is analyzed for the use of vocabulary that could indicate a specific value. For this purpose, a dictionary specifically developed to assess references to personal values [47] was utilized, and points were awarded for each of the values depending on the detected words. Using these points, the network updates the conditional probability of the chosen value, given the specific context.

When the agent has completed the updates, it can calculate the conditional probabilities of each value and, depending on which ones are more prevalent, can choose the corresponding scenarios to discuss further and get more information on. The ultimate use of this system emerges in the second session of the experiment where the two memories work in tandem. The robot uses semantic memory to make a general statement about the participant and their values, then substantiates this generalization with an example from the episodic memory.

3.4 Follow-up questions

For the semantic memory to achieve a value profile as close to the participant as possible, the Bayesian Network needs to be updated with a sufficient amount of data. Since a minimal combination of values and situational characteristics are tested with the scenarios, there needs to be a way to enhance the network to more accurately calculate the value profile and reduce uncertainty. Since introducing new scenarios would be too time-consuming, a change to the original scenarios is proposed. In the second session, after discussing the importance and consequences of their decision, the robot suggests a slight alteration in a way that changes the situational characteristics of the specific scenario, while the options,

#	**Original D.I.A.**	**Follow-up D.I.A.**	**Possible Alteration**
1	0,0,1	1,0,1	she had done the same to others before, but they did not speak up
2	0,1,0	0,1,1	it was a really good friend of yours and would accuse you of being a bad friend
3	0,1,1	0,1,0	there were more people in the group that did not agree with that topic
4	1,1,1	1,0,1	it was just a normal class, but your professor strongly disagreed with the strike

Table 3.2: Examples of follow-ups previous scenarios. Situational characteristics (Duty, Adversity, Intellect) are denoted with 1 if present and 0 if absent.

and therefore the value conflict, remain the same. The alteration depends on the previous answer of the subject, while subtly attempting to push them towards the other option. The Bayesian Network is updated with this new information and the assumption is that the conditional probabilities will be more in line with the participant's behavior. The proposed alterations for the example scenarios of Table 3.1 are given in Table 3.2.

3.5 Feedback

As described in Section 2.2.3, the feedback provided to the participants is meant to motivate them and increase their feeling of competence so they're more engaged during the conversation. Feedback during the interaction is given in 2 ways. Firstly, when the subject is asked to justify their decisions if their argument is not sufficiently explained, the robot inquires further about their reasoning, selecting from a pool of more in-depth questions, which are designed to extract more information from the participant, while stating that what they just provided is not adequate (e.g. "It's not quite clear to me yet why you would do this, can you talk a little more about it?"). If they fail to do so for a second time, an encouraging but critical prompt is given (e.g. "Maybe think on it for a little longer and we'll talk about it next time."). Furthermore, after an argument has been provided, the robot gives positive non-generic feedback, praising the quality of the explanation but not the choice itself (e.g. "Great explanation!"). Another great form of feedback that this particular study would have benefited from is gestures. However, the limited capabilities of the robot, in combination with the noises produced during its movements, would have impacted the user's experience and the data collection greatly and therefore gestures were not included in the final design.

3.6 Robot manipulation

A major issue that was apparent during previous studies is the speech recognition of robots, and more specifically the NAO robot[48, 49, 50]. The noises produced by the robot itself, the distance from the participant that is talking, the participant's pitch and volume, and noises from the environment, are all things that can severely affect the speech recognition software [51] and reduce the quality of the interaction, and subsequently the participant's engagement. For this reason, a decision was made to manipulate the robot through a Wizard of Oz(WoZ) interface. The robot's dialogue was scripted and programmed into the application, but the participant's answers were provided by the experimenter by choosing between buttons on a different program, unbeknownst to the participant, to maintain the illusion of an actual conversation with the robot. The WoZ program was designed to only register answers to multiple-choice questions or advance the dialogue since the delay in typing an answer to an open-ended question would give away the pretense.

3.7 Argument Quality

For the ultimate goal of the experiment, which is to enhance the participant's understanding of values, it is important to set learning goals in the beginning and define what the participants should gain from the interaction. To measure this aspect, the quality of argumentation is introduced and the goal is to improve the quality over the two sessions.

Self-regulated learning (SRL) refers to the proactive process of taking control of the learning experience and can facilitate the setting of goals. It involves assessing oneself, setting goals, employing appropriate strategies, and monitoring one's progress with the intention of academic advancement[52]. Many SRL models have been developed over the years [53], but the one referenced in this thesis is the cyclical phase model Zimmerman and Moylan adapted[54] (Figure 3.4).

SRL is closely intertwined with motivation and can have a reciprocal effect. On the one hand, it can enhance and maintain motivation through self-motivation beliefs. On the other hand, when motivation is high, the effects of SRL can be more impactful.

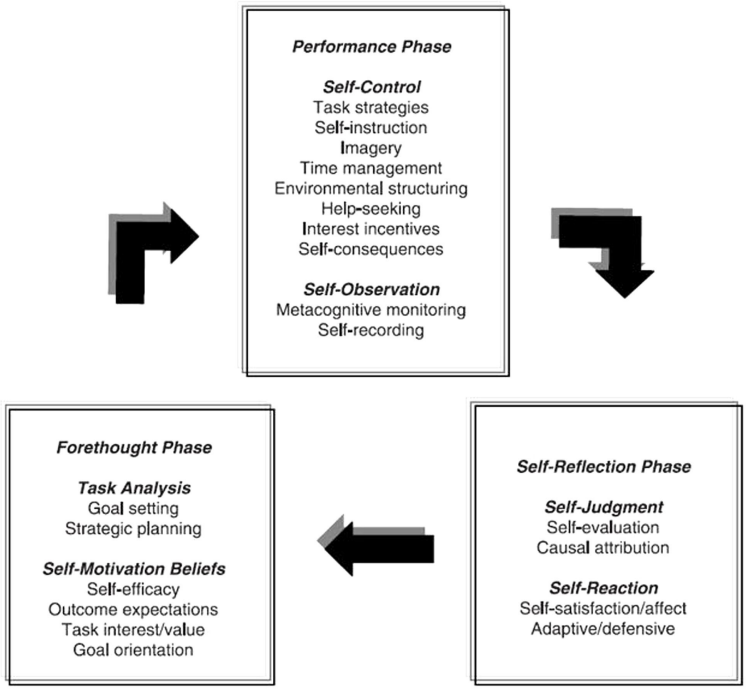

Figure 3.4: SRL model adapted from Zimmerman and Moyle[54]

To assess the quality of the participants' arguments, multiple schemes were considered. The first scheme considered was the Structured of Observed Learning Outcomes ($SOLO$) taxonomy[55], which categorizes the arguments into five distinct levels: prestructural, unistructural, multistructural, relational, and extended abstract levels. Despite the comprehensive nature of this taxonomy, the nature of the questions does not allow for answers of sufficient length to cover the whole spectrum and, therefore there is a need for an alternative scheme. Given that the objective of the experiment is to elicit insightful arguments that analyze the reasons for certain behaviors, it was important to consider both categories of arguments, as well as levels of arguments. The combination of the two has been used before, as documented by Ullmann[56], and it entails a two-stage coding system that addresses the two dimensions, depth (categories) and breadth (levels) is widely used to map both qualities.

For the first dimension, two categories were chosen: descriptive and reflective. An argument is characterized as descriptive if it provides a description of the current situation, or of the possible outcomes given the behavior choices. Respectively, an argument is characterized

Depth Dimension			**Breadth Dimension**		
Categories	Explanation	Value	Levels	Explanation	Value
Reflective	thoughts going through their mind, expressing likes and beliefs(values)	1.0	Report	no mention of any value	-1.0
Descriptive	description of the situation in the real world, or possible outcomes of the options	0.0	Mention	brief or implied mention of a value	-0.5
			Express	direct or implied expression of a value	0.5
			Analyze	in-depth expression and analysis of a value	1.0

Table 3.3: 2-stage argument quality assessment scheme

as reflective if it includes the thoughts going through the participant's mind, not limited to describing the situation, or if it expresses their likes and beliefs. For the second category, there needed to be distinct levels of value inclusion in the provided arguments. Isaacs et. al [57] define a similar scheme in their paper regarding the emotional depth of participants' reflections. Their scheme regards four mutually exclusive levels: Report, Mention, Express, and Analyze. For the purposes of this study, this scheme has been altered to reflect the depth of value reference. An argument that made no mention, explicit or implicit of any value, is considered to be of level "Report". If there is a brief mention of a value, that argument is considered to be in the "Mention" level. However, if there is a direct or implied expression of a value, that argument belongs in the "Express" level. Subsequently, the "Analyze" level is reserved for arguments that include an in-depth expression and analysis of a value.

Since this study aims to elicit thought and reflection, a reflective argument is considered better than a descriptive one, and respectively, the Breadth levels can be considered ordinal, with "Report" being the lowest one, and "Analyze" being the highest. Therefore, they can be quantified and mapped to a value. The values associated with each level, as well as an overview of the assessment scheme for argument quality, can be found in Table 3.3.

Chapter 4

Method

4.1 Participants

A total of fifty-five adults participated in the experiment, ages between 22 and 38. The initial participants were fifty-seven, but two of them did not complete the second session within the decided time window and were therefore excluded from the study. The number of subjects that identified as male is significantly larger (37) than female (18). The participants were recruited at the TU Delft campus and were mostly students, former students, or employees of the university. Two of the subjects have completed their high school education, thirteen have attained their Bachelor's diploma, thirty-seven have Master's level education, and three were at a level of Ph.D. or higher. The participants were separated into two conditions randomly and of the 55, 28 were in the baseline condition, whereas 27 were in the memory condition.

4.2 Material and Setup

Throughout all interactions, the setup remained consistent (Figure 4.1). The robot that was used is the NAO V6 of Interactive Robotics and it was positioned on a desk while the participant sat in a chair facing towards it. Between them, there was a screen that displayed the images relevant to each scenario. A microphone was placed in front of them, next to the screen, and connected to the experimenter's laptop to record the conversation and the subsequent interview with the experimenter. The experimenter was seated to the back left

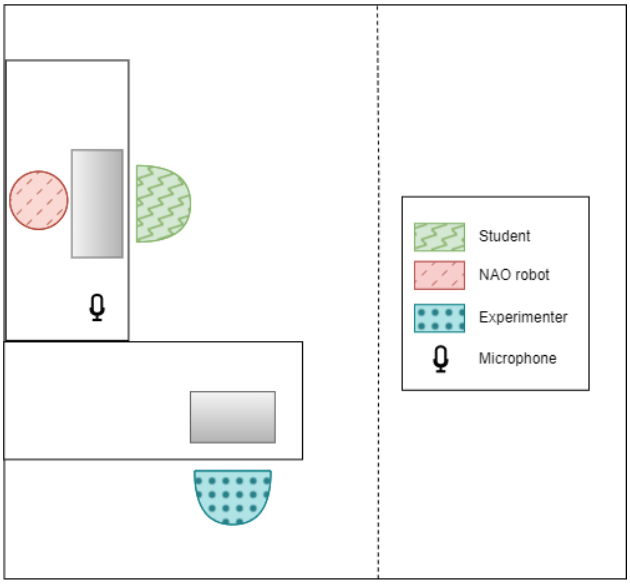

Figure 4.1: Study Set-up

of the participant with a Dell XPS laptop which was used to run the application. The audio was recorded through the Audacity program.

4.3 Questionnaires

4.3.1 Pre-Session 1 Questionnaire

Throughout the study, participants are asked to complete two questionnaires: one before the first session, referred to as the pre-session 1 questionnaire, and one after the second session, termed the post-session 2 questionnaire.

The pre-session 1 questionnaire includes some demographic questions, a subset of the Revision of the Self-Monitoring Scale, and a ranking of their personal values.

Self-Monitoring Scale

The Self-Monitoring Scale, initially developed by Snyder[58], and later revised by Lennox and Wolfe[59], is used to investigate consistency in expression across situations and would indicate which participants are more in control of their reactions. The Revised Self-Monitoring

Scale is comprised of two subscales, "Ability to modify self-presentation" and "Sensitivity to expressive behavior of others", which can be considered together, as well as separately. In this study, the most relevant of the two is the ability to modify self-presentation and thus is the one included in the questionnaire. The items of the subscale are as follows:

1. In social situations, I have the ability to alter my behavior if I feel that something else is called for.
2. I have the ability to control the way I come across to people, depending on the impression I wish to give them.
3. When I feel that the image I am portraying isn't working, I can readily change it to something that does.
4. I have trouble changing my behavior to suit different people and different situations.
5. I have found that I can adjust my behavior to meet the requirements of any situation I find myself in.
6. Even when it might be to my advantage, I have difficulty putting up a good front.
7. Once I know what the situation calls for, it's easy for me to regulate my actions accordingly.

The above items are measured with a 6-point Likert scale represented by levels of agreement with the statements. The choices available to the participants were: 'Always true', 'Generally true', 'Somewhat true', 'Somewhat false', 'Generally false', and 'Always false'.

Personal Value Ranking

Participants are also tasked with ranking six of the Schwartz personal values based on their perceived importance. In addition to "Benevolence", "Achievement", "Self-Direction", and "Conformity", the values included in the list are "Power" and "Hedonism". These values are provided as a distraction so that the participants are not biased when responding to the scenarios later on, as well as to have a better estimation of the relative distance between the

Concept	Items
Anthropomorphism	Fake – Natural Machinelike – Humanlike Artificial – Lifelike
Animacy	Stagnant – Lively Apathetic – Responsive Inert – Interactive
Likeability	Unfriendly – Friendly Awful – Nice
Perceived Intelligence	Incompetent – Competent Foolish –Sensible

Table 4.1: Subset of items from the Godspeed Questionnaire included in the post-session 2 questionnaire.

considered values. The participants were given papers representing each value and, initially starting with two, were asked to place them in a descending order of importance. A new value was introduced at every step until all values were ranked. In the same questionnaire, the participants were also asked about their confidence in the ranking, with a 5-point Likert scale. The same variable is gathered in the post-session 2 questionnaire, in the same manner.

4.3.2 Post-Session 2 Questionnaire

The post-session 2 questionnaire includes a subset of the Godspeed questionnaire, a choice between two value profiles in the form of radar plots, and, once again, a personal ranking of a subset of the Schwartz values.

Godspeed Questionnaire

The Godspeed Questionnaire [60] was developed as a standardized measurement tool for human-robot interactions and measures 5 concepts: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Each of these concepts is evaluated through different items using a semantic differential scale that ranges from 1 to 5. A total of 10 items were selected for this study, comprising 3 each from the "anthropomorphism" and "animacy" concepts, and 2 each from the "likeability" and "perceived intelligence" concepts. No items were included from the "perceived safety" concept as it was deemed unrelated to the topic. The full list of items in the questionnaire can be found in Table 4.1.

Value Profile Plots

At the end of the experiment, the participants are asked to choose between 2 value profiles, meaning the robot's estimation of how important they consider each of the 4 values. One of the profiles constitutes the true probabilities estimated by the Bayesian network, while the second is initialized as such but then a random number between 0.8 and 1.5 is added or subtracted from each of the original values. The participants are shown an image that displays both value profiles on the same radar plot, the function of a radar plot is explained to them, and then they are given time to look at and interact with the plot to make their decision. After choosing the one which they think better represents them and how important they find these values, they are asked to give a confidence score for their choice.

4.4 Procedure

4.4.1 Session 1

Before the first session, the participants are asked to read and sign the consent form which is also explained to them verbally. At this point, they also fill in the first questionnaire, and the researcher keeps track of their value ranking to input into the system.

In the next stage, the subject takes their place in front of the robot, the experimenter enters the ranking of the values into the program and initiates the audio recording, the WoZ interface, and the robot application. In session 1, the interaction is almost identical for the two conditions (baseline/memory). First, the robot provides a brief introduction, introducing itself and explaining the purpose of their conversation. For each of the six scenarios, the robot describes the situation and the different behavior options, while it shows the corresponding image on the screen in front of it. In the case of the memory condition, at this moment, the robot verbally cues toward the screen so that the participant pays closer attention to the images. In the baseline condition, no such cue is given. The robot then asks for the participant's behavior choice in that situation, the reasoning behind their answer, as well as two questions regarding how difficult it was to make the decision and if they have previously been through a similar situation. When all the scenarios have been discussed, a

brief outro is presented, marking the conclusion of the interaction. After the first session, the experimenter interviews the subjects regarding their recollections of the scenarios discussed. The full process is described later in this section (4.4.4).

4.4.2 In-Between Sessions

Between session 1 and session 2, the experimenter goes through the audio files and records the given answers which will be used in session 2. These are restructured to be more clear and concise.

The goal was to keep the time that transpires between the two sessions around 7 days, with the final average calculated at 8.164 days. According to Ebbinghaus's Curve of Forgetting[61], which refers to the retaining of newly learned information during learning, humans will forget almost 75% of the acquired knowledge within 2 days. Since Ebbinghaus's study attempted the retention of nonsensical syllables which can be hard to remember, in this study a longer time window is considered to account for the more meaningful information, that although new, should be easier to remember within the context.

4.4.3 Session 2

For session 2, the difference between the two conditions is more prominent and should be described separately. For the baseline condition, after a short introduction where the robot does not show any signs of recognizing its conversation partner, three scenarios are picked randomly, from those that were not discussed during session 1. For these scenarios, the order of events is similar to session 1. A short description is given, followed by the choice sequence, where the subject has to decide and justify their choice, as well as answer about the difficulty and the relatability of the scenario. At this point, the concepts of values and context are introduced. The participant is asked to self-reflect on why they consider their choice important, and what its consequences are, and finally, the robot asks the follow-up question. When the robot has discussed all three scenarios, a small outro is given which provides a quick summary of the interactions.

For the memory condition, the robot greets the participant by name and introduces

the topic of this interaction. The three values with the highest posterior probabilities are extracted from the Bayesian Network of the semantic memory, and, for each value, one scenario discussed in session 1 where that value was chosen is selected to be discussed further. The robot reminds the participant of the scenario, cues towards the images on the screen, and then asks if they remember what they had answered in the previous conversation. The robot responds accordingly, reminds the participant of their argument, and, after introducing the values and the contexts, asks the subject to self-reflect on why that choice was important to them. Similarly to the baseline condition, they then move on to talk about the consequences of their choices, and a follow-up question about an alternative context, while the interaction ends with a short conclusion. The interview regarding recollections is repeated for this interaction, similar to session 1.

After the second interaction, the post-session 2 questionnaire is given to the participants.

4.4.4 Post-Interaction Interview

Immediately after each interaction, the experimenter conducts a short interview to assess the quality of the recollections and the memorability of the scenarios. This factor indicates the participants' attentiveness during the conversation, and various reasons may account for the differences in recollections among individuals. The interview follows a semi-scripted format where specific questions are posed to all participants, but the experimenter is free to seek additional information depending on the responses and the expressiveness of each subject. The general script is adapted from the one crafted by Nikkels[46] for Burger's study, and the two-part format, which was optional, was established for all individuals. First, the participant is questioned on memories of the interaction that they can freely recall, and second, the experimenter presents the images that appeared on the screen during the scenarios and inquires further about them. The script that was followed can be found in Appendix B.

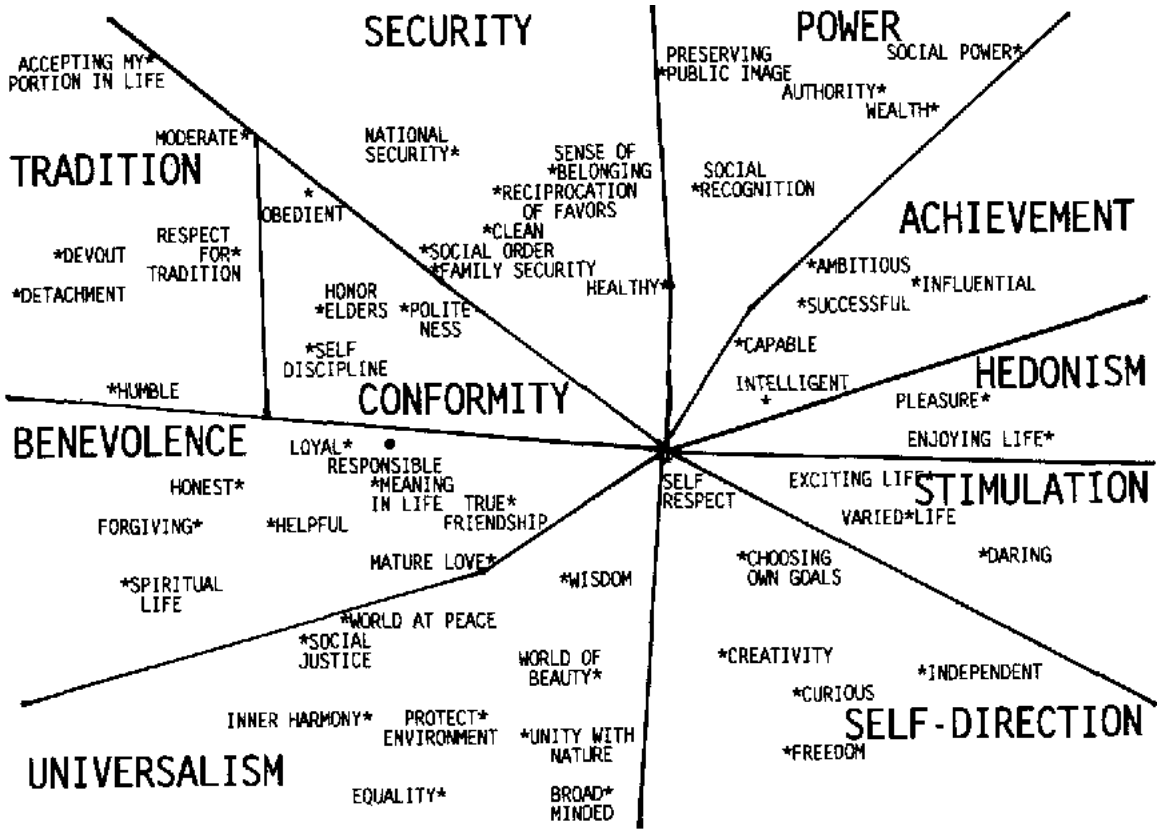

Figure 4.2: Subvalue Structure for Value Annotation

4.5 Variables

In the following section, different variables collected during the experiment or the processing of the data are described in detail, before moving to the Data Analysis chapter.

4.5.1 Argument Annotation

Each of the arguments provided by the participants was annotated according to the scheme described in Section 3.7 for argument quality and therefore are associated with a category of "Depth" and a level of "Breadth". Moreover, the argument is examined for the value it expresses, according to the Schwartz circumplex model and the subvalue structure[13] (Figure 4.2), and therefore has an "Argument Value" label.

During the first session, one argument was collected per scenario discussed, with a total of

6 scenarios per participant. For the second session, in the baseline condition, two arguments were collected for each scenario, while in the memory only one, with a total of 3 scenarios discussed. This is due to the fact that, in the baseline condition, a new scenario was introduced, so the robot asked for the reason for choosing a behavior ('choice' argument). In both conditions, the subjects are questioned on whether they consider a particular value important based on the robot's assumptions, and the reason for their perspective('importance' argument). The information about their agreement or disagreement with that assumption is also recorded and shall now on be mentioned as "Assumption Perception".

Additionally, each argument has an "Assumed Value" based on the value associated with the chosen behavior. To verify that this assumption made by the robot is correct, the "Assumed Value" is compared to the "Argument Value" and it is determined if these two values match. Two methods of comparison are introduced and will be studied in the Data Analysis Section. The first is the "Exact Match", meaning that the values match exactly. However, the Schwartz model is circular and many of the values can be close to two or more others. Since this study only examines 4 of the 10 Schwartz values, it is only understandable that it cannot cover all the values displayed by the participants, and it is deemed beneficial to perform a "Close Match", meaning matching the two variables in the case that they are next to each other in the Schwartz circumplex model.

In summation, each argument is characterized by 6 (or 7) annotations: Depth, Breadth, Argument Value, Assumed Value, Exact Match, Close Match, and Assumption Perception in the case of the 'importance' arguments.

4.5.2 Recollections Quality

The recollections gathered during the post-interaction interviews were categorized into 3 levels, each of incremental quality, for every scenario mentioned by the participant. These annotations were derived from the experimenter's notes taken during the interview and are defined as follows: at level 'Low', the participant could recall the scenario and its details, at level 'Medium', the participant could also describe the images displayed on the screen with some inaccuracies, and at level 'High', the participant could accurately recollect the presented

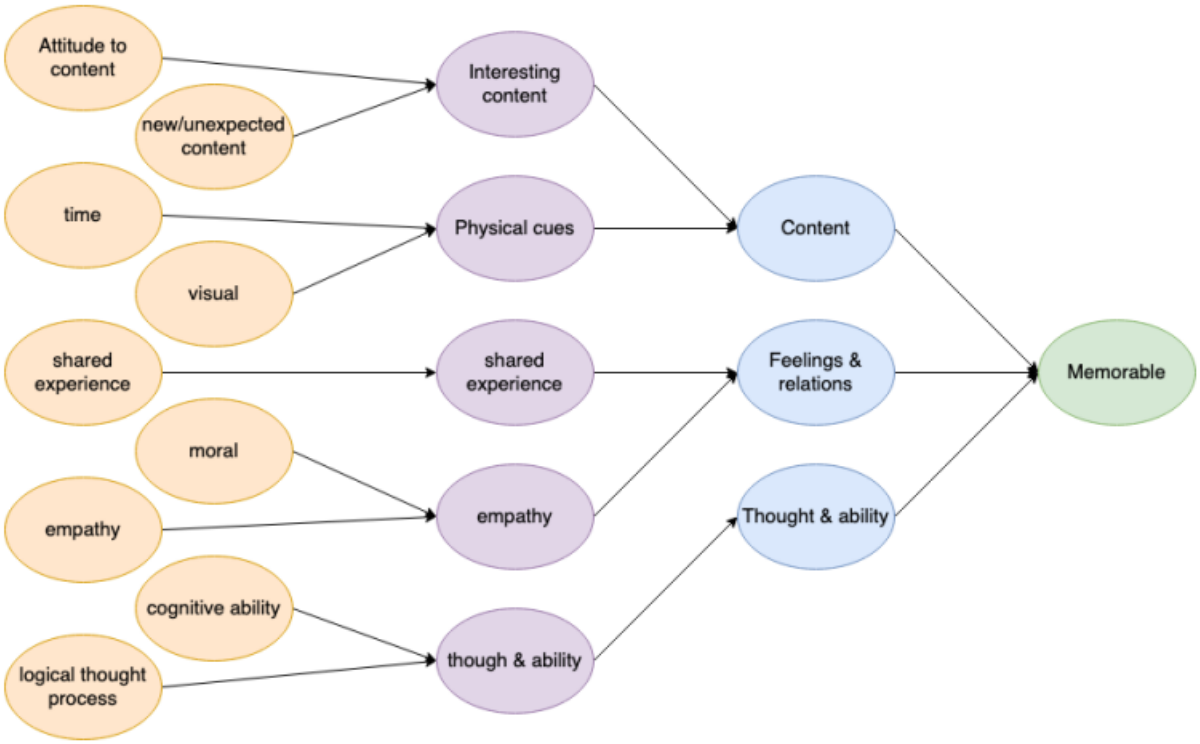

Figure 4.3: Labeling Hierarchy for reasons for remembering

images along with the scenario details. For better analysis of the data, a quantitative scale from 1 to 3 is defined for the three possible levels.

4.5.3 Reasons for Remembering

Regarding the reasons for remembering particular scenarios, due to the shared topic of conversation, the annotations are based on the analysis done by Nikkels[46], which is also loosely based on the work of Tsfasman[62]. Nevertheless, these annotations have been tailored to the specific themes inherent in the dataset. The resultant thematic analysis is presented in Figure 4.3.

4.5.4 Plot Choice

For each participant, the information about which Value Profile Plot they chose, and with what confidence, will be important for the analysis. The plot choice is a binary variable with values 'Robot's Plot' and 'Random Plot', while the confidence is measured in a 5-point Likert scale where 1 is *Not Confident*, and 5 is *Very Confident*.

4.5.5 Godspeed Questionnaire results

The recorded answers to the Godspeed questionnaire are also of importance for Hypothesis 4. Each of the questions is answered on a 5-point Likert scale and, for each concept, all questions are considered of equal weight. Therefore, the average for each concept will be examined instead.

Chapter 5

Data Analysis

In this section, a thorough explanation of the processing and analysis of the data collected during the experiment will be provided. Subsequently, the results will be interpreted in an attempt to prove the hypotheses presented in section 2.4.

5.1 Data Processing

5.1.1 Interrater reliability

Since a large portion of the variables consist of manually annotated data, an interrater reliability test is required to establish the validity of the annotations. This regards the variables of the arguments, meaning their Depth, Breadth, and Value. The total amount of the arguments were annotated by the main researcher, while 15% of them were annotated by a second annotator who was given the argument quality scheme and instructions on how to assess the arguments. Krippendorff's alpha was used to evaluate the inter-annotator agreement. For the Depth variable, using an ordinal kernel, the value achieved was $\alpha = 0.819$, for the Breadth variable, an ordinal kernel was used which resulted in $\alpha = 0.846$, while for the Argument Value variable, a circular kernel was applied, to account for the Schwartz circumplex model, and produced a value of $\alpha = 0.777$. We can therefore conclude that for the Depth and Breadth variables, the agreement was reliable ($\alpha > 0.8$), whereas for the Argument's Value variable, tentative conclusions can be made ($\alpha > 0.667$).

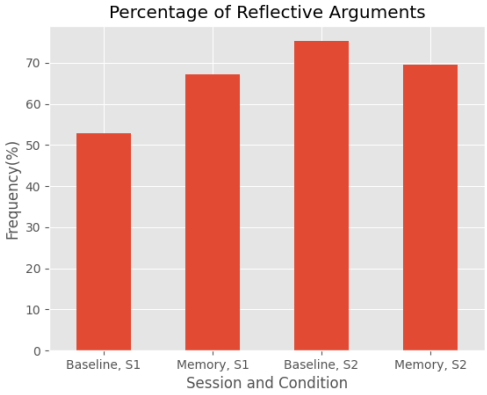

Figure 5.1: Percentage of Reflective Arguments per Condition per Session

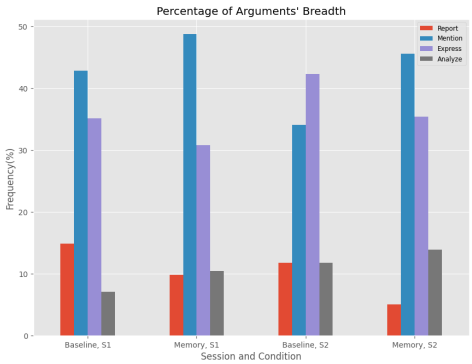

Figure 5.2: Frequency of Breadth Level Values per Condition per Session

5.1.2 Incomplete data

Two of the total subjects in the study participated only in the first session of the experiment. The data from those sessions was disregarded and is not included in the analysis.

5.2 Statistical Analysis

5.2.1 Hypothesis 1

Participants in the memory condition will have a better quality of argumentation, especially in the second interaction.

The quality of argumentation is defined by a 2-level scheme as delineated in section 3.7. For the Depth variable, only the reflective arguments will be reported as the two values are complementary. In Figures 5.1 and 5.2, the frequency of the values for the Depth and Breadth variables are presented.

Firstly, the correlation between the two variables needs to be examined. The Shapiro-Wilk test reveals that both variables are non-normally distributed (Depth: $W = 0.61, p < 2.2e-16$, Breadth: $W = 0.87, p < 2.2e-16$). The more fitting correlation test is Spearman's correlation coefficient which weakly relates the two variables ($r = 0.28, p = 4.325e-10$), therefore they can be explored independently.

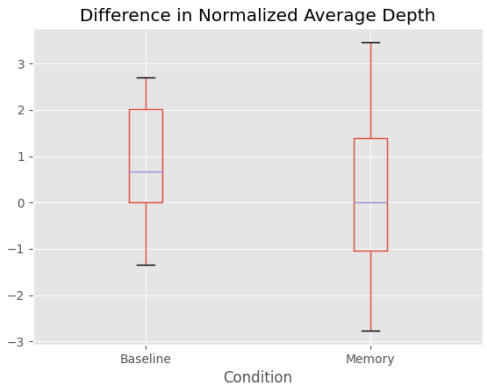

Figure 5.3: Distribution of Difference of Normalized Averaged Depth Level per Condition

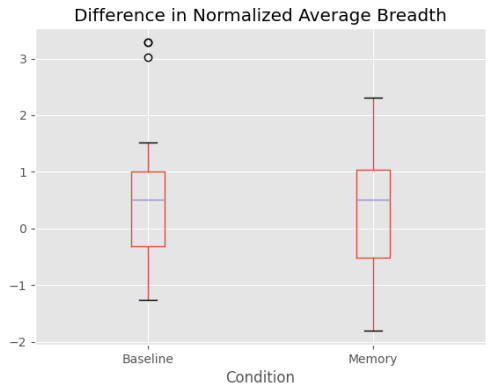

Figure 5.4: Distribution of Difference of Normalized Averaged Breadth Level per Condition

In Figure 5.1, a noticeable difference can be observed for session 1 between the two conditions. To further examine the effect, we average the number of reflective arguments per participant, and the ANOVA test reveals a significant difference ($F(1, 53) = 6.62, p = 0.0129, r = 0.673$) between the baseline and the memory condition with a large effect size. Therefore, even from the first interaction, participants in the memory condition seem to reflect more in their arguments.

Since the interest is in the improvement of the argumentation quality over the 2 sessions, the difference between the 2 sessions is considered. First, both Depth and Breadth levels are normalized per condition using z-score to account for the differences between participants in the conditions. Then, for each participant, the Depth and Breadth levels are averaged per session and their difference is calculated. After confirming that the data follow a normal distribution (Depth: $W = 0.97, p = 0.275$, Breadth: $W = 0.96, p = 0.067$), and are homogeneous for both conditions (Depth: $F(1, 53) = 0.68, p = 0.413$, Breadth: $F(1, 53) = 0.04, p = 0.842$), a one-way ANOVA test was performed which showed a significant effect between Depth progression and condition with large effect size ($F(1, 53) = 4.83, p = 0.0324, r = 0.553$). No significant effect was found between Breadth progression and condition ($F(1, 53) = 0.054, p = 0.817, r = 0.007$). The distribution of the differences in average levels for Depth and Breadth can be found in Figures 5.3 and 5.4.

The difference for Depth in the baseline condition seems to be larger, looking at figures 5.1 and 5.3, but that is because reflective arguments were already more frequent for the memory condition in session 1, so there was no room for much improvement.

About the arguments' Breadth, it can be said from Figure 5.2 that 'Report' arguments seem to decrease while 'Analyze' arguments seem to increase for both conditions, however, overall, no significance can be attributed to this effect.

5.2.2 Hypothesis 2

Participants in the memory condition are more likely to agree with the robot's value profile about them.

To test this hypothesis, the variables for Plot Choice and Confidence will be utilized. Figure 5.5 shows which plot participants identified as more representative of their own values for each condition. The number of participants that correctly identified the robot's plot is the same for both conditions, while the only difference in the random plot is because of the uneven number of subjects in the conditions. This means that the hypothesis cannot be confirmed, however, it is still interesting to study to try and understand why this happens.

Looking at the frequency of the confidence scores (Figure 5.6), the participants in the memory condition seem to overall rate their confidence in choosing the correct plot higher. The Shapiro-Wilk normality test showed that the distribution for confidence is not normal (baseline: $p = 0.0014$, $W = 0.858$, memory: $p = 0.0006$, $W = 0.836$). The Wilcoxon rank-sum test shows that the confidence of choice between the baseline condition group($Mdn = 4.0$) and the memory condition group($Mdn = 4.0$) does not differ significantly ($W = 284.5, p = 0.0916, r = -0.228$), while the difference of confidence between participants that chose the robot's plot ($Mdn = 4.0$) and participants that chose the random plot ($Mdn = 4.0$) is also not significant ($W = 384, p = 0.704, r = -0.0512$). Even if there is no significance to the data, in Figure 5.7 there is a clear trend of the memory condition group that chose the random plot to be the most confident of their choice.

The reason for this discrepancy should be more clear after the next hypothesis exploration.

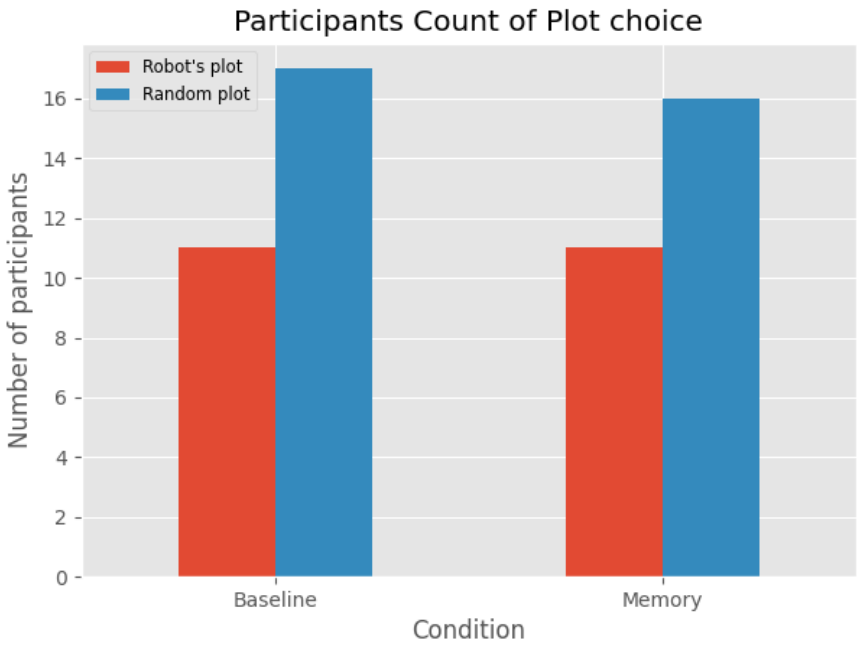

Figure 5.5: Number of Participants per Plot Choice per Condition

5.2.3 Hypothesis 3

Participants whose arguments reflect the assumed value of their choice are more likely to identify the robot's value profile correctly.

In the previous hypothesis, although it was not possible to confirm it, an interesting discrepancy was identified and the reason for it could be linked to the results of this hypothesis. Since the robot's model only includes 4 of the 10 Schwartz values, the assumptions that the robot makes are not going to always be correct. Human motivation is multi-faceted and cannot always be limited to one reason. Therefore, the participants in the memory condition that chose the random plot might understand their values much more, but the robot might not have been as good at assessing them. In this section, the link between plot choice and the assumptions of the robot will be explored.

For each participant, a percentage of matching values was determined from the overall arguments provided. For the two variables of 'Exact Match' and 'Close Match', the distribution of those percentages per Plot Choice can be seen in Figures 5.8 and 5.9.

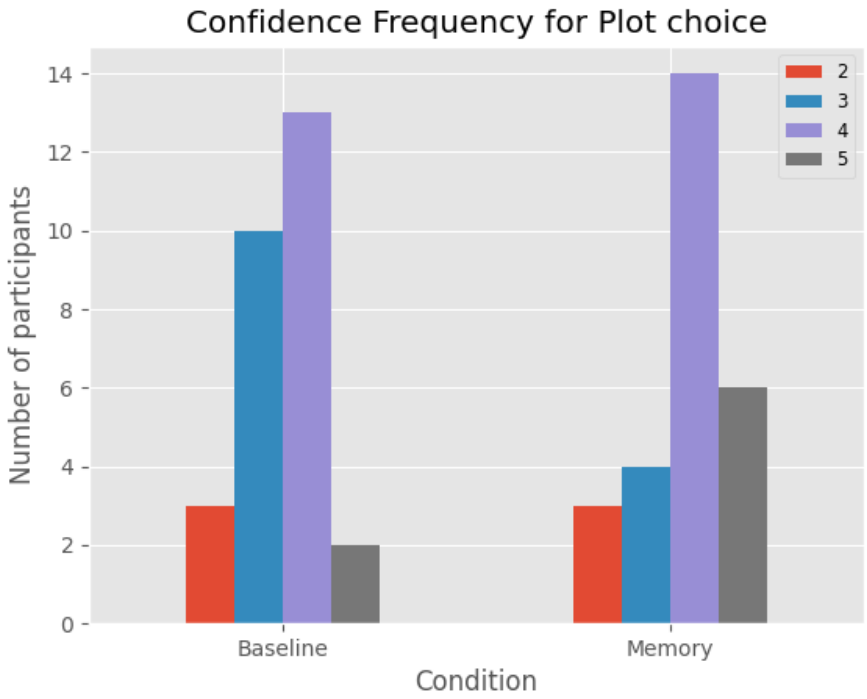

Figure 5.6: Frequency of Confidence values per Condition

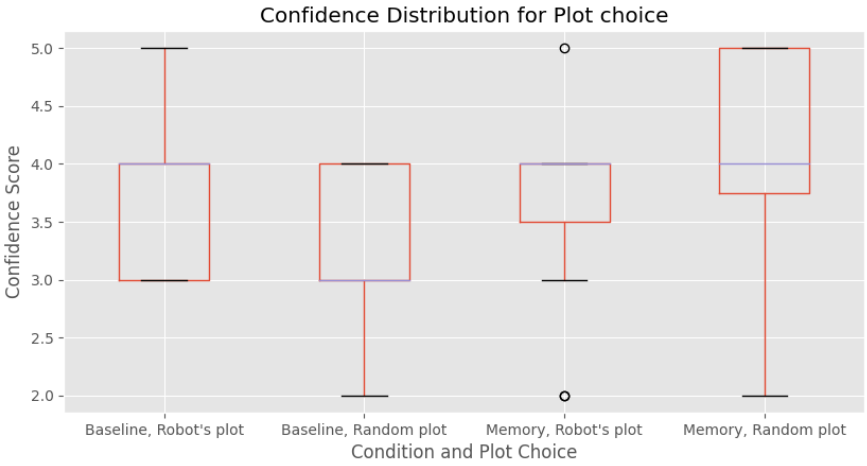

Figure 5.7: Distribution of Confidence for participants in the two conditions per Plot Choice

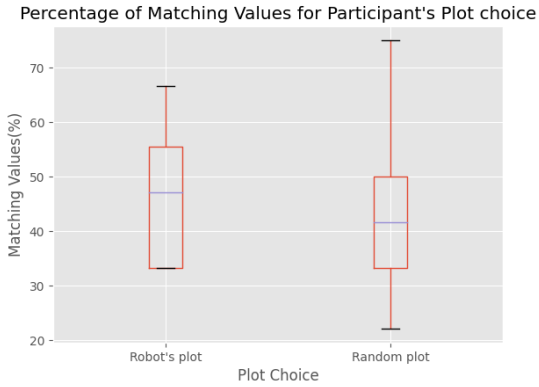

Figure 5.8: Exact Match Percentage distribution per Plot Choice

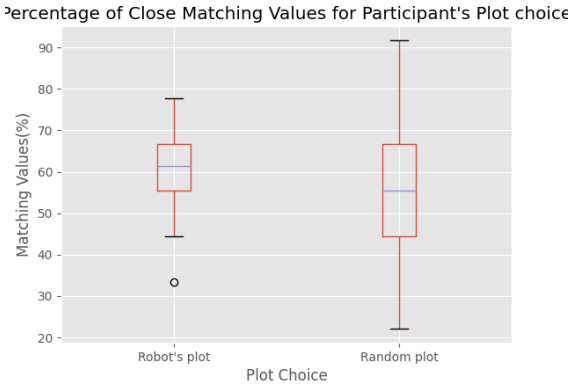

Figure 5.9: Close Match Percentage distribution per Plot Choice

It appears that the overall percentages of the Close Match tend to be higher($Mdn = 58.33$) than those of the Exact Match($Mdn = 44.44$). Both variables don't follow a normal distribution according to the Shapiro-Wilk normality test (Exact: $W = 0.96, p = 0.043$, Close: $W = 0.95, p = 0.02$), so the Wilcoxon rank-sum test is performed to compare between the different plot choices. As is apparent from Figures 5.8 and 5.9, no significant difference was found between participants that chose the Robot's plot (Exact: $Mdn = 47.22$, Close: $Mdn = 61.31$) and those that chose the Random Plot (Exact: $Mdn = 41.67$, Close: $Mdn = 55.56$) for either Exact Match ($W = 450, p = 0.132, r = -0.203$) or Close Match ($W = 465.5, p = 0.077, r = -0.238$). Close Match is slightly more significant and also is a better representation of value relations, so for the rest of the analysis, when discussing matching argument values, it is assumed that they are closely matching.

At this point, it is important to explore the self-reported measure of Assumption Perception, which denotes how participants perceived the assumptions made by the robot. In Figure 5.10, the frequency of each of the values is displayed per Plot Choice. Participants who chose the robot's plot appear to have answered "sometimes" three times more often than those who chose the random plot (22,73% vs 7.14%), which could indicate that they understand the intricacies of decision-making, and how context can play an important role. It is also important to note that participants who chose the random plot disagreed two times

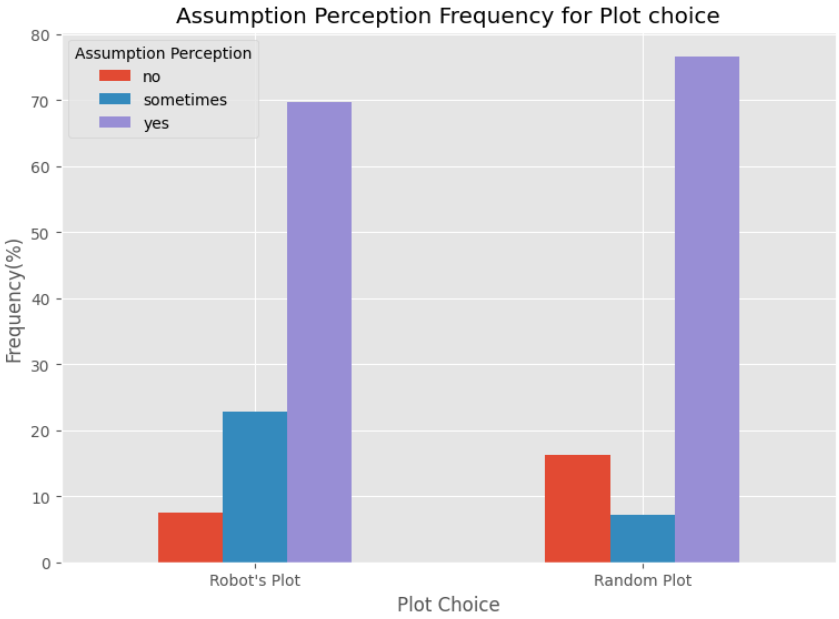

Figure 5.10: Assumption Perception Frequency per Plot Choice

more often with the robot's assumption (7.58% vs 16.33%). That is aligned with the theory that the model for the values was not able to fully capture the subjects' value profile.

To further investigate this theory, the relation between matching values and assumption agreement is explored. In Figure 5.11, the percentages of agreement (both 'yes' and 'sometimes') are considered depending on which plot was chosen and if the values of the arguments match. Although not significant, the frequency of agreement is lower for those who chose the random plot. Additionally, between arguments in which the values matched and those that didn't, for both plot choices, the agreement was lower, which is expected.

5.2.4 Hypothesis 4

Participants in the memory condition will find the robot more likable and lifelike.

For this hypothesis, we will utilize the results of the Godspeed questionnaire. The distribution of the ratings per question is shown in Figure 5.12, and per section of the questionnaire in Figure 5.13. The question ratings for each section have similar distributions, while overall,

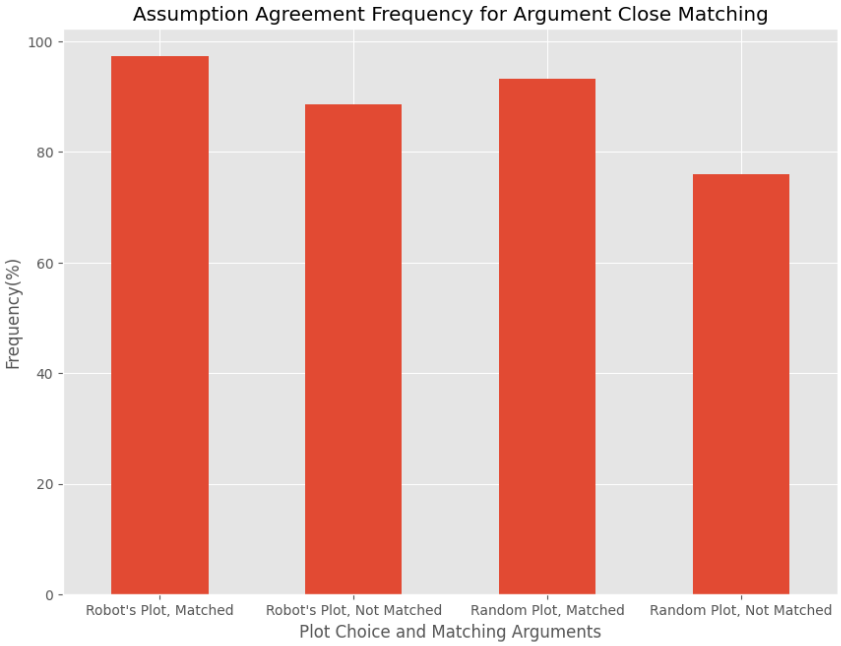

Figure 5.11: Assumption Agreement frequency for Plot Choice and Value Matching

'Likability' is rated the highest, and Anthropomorphism the lowest.

Figure 5.14 shows the ratings for each section per condition. A one-way independent t-test proves that none of the sections are significantly higher in the memory condition, with Likability even being slightly lower (Anthropomorphism: $t(53) = 0.375, p = 0.355$, Animacy: $t(53) = 0.865, p = 0.195$,Likability: $t(53) = -0.227, p = 0.589$, Perceived Intelligence: $t(53) = 1.403, p = 0.0832$).

5.2.5 Hypothesis 5

Participants in the memory condition will have better recollection quality than those in the baseline condition.

First, the average number of scenarios that participants remembered is explored for this hypothesis. In Figure 5.15, the boxplot shows the number of scenarios that participants remembered in session 1 for the baseline ($M = 4.25, SD = 1.076$) and for the memory

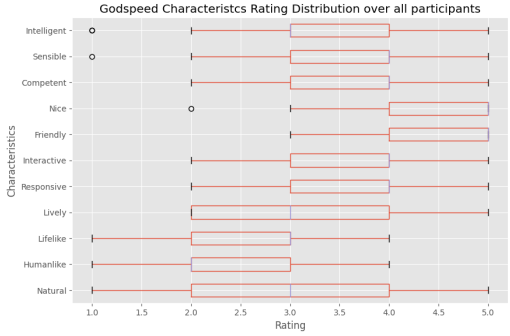

Figure 5.12: Godspeed Questionnaire Ratings Distribution per Question

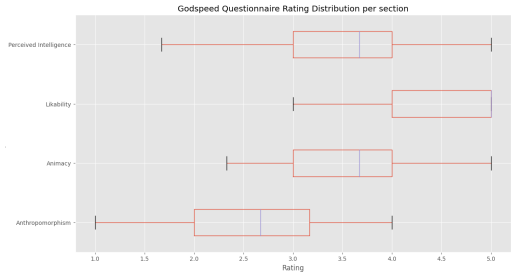

Figure 5.13: Godspeed Questionnaire Ratings Distribution per Section

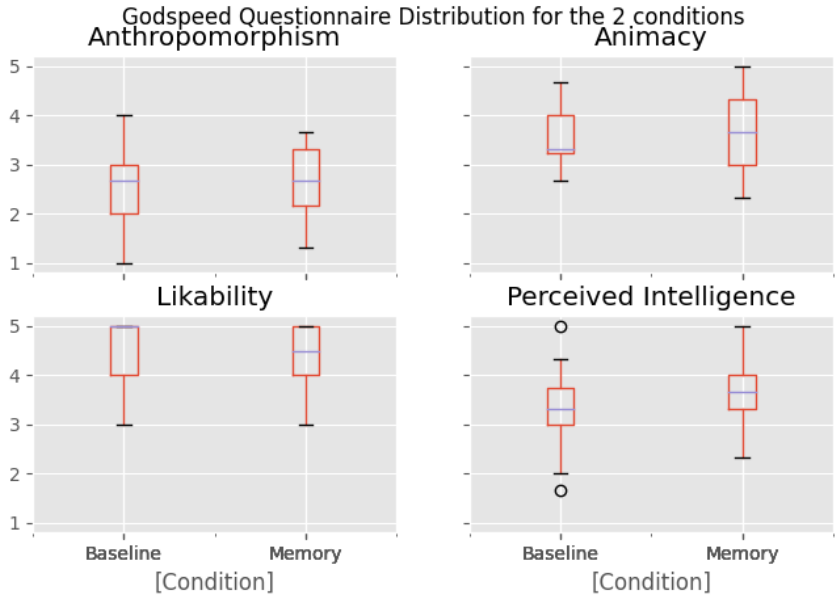

Figure 5.14: Godspeed Questionnaire Ratings per Condition per Section

condition ($M = 4.0, SD = 1.24$), as well as in session 2(baseline:$M = 2.71, SD = 0.535$, memory: $M = 2.89, SD = 0.32$). Since in session 2, only three scenarios were discussed, which was reportedly easier to remember, it is expected that the means are closer to the total number.

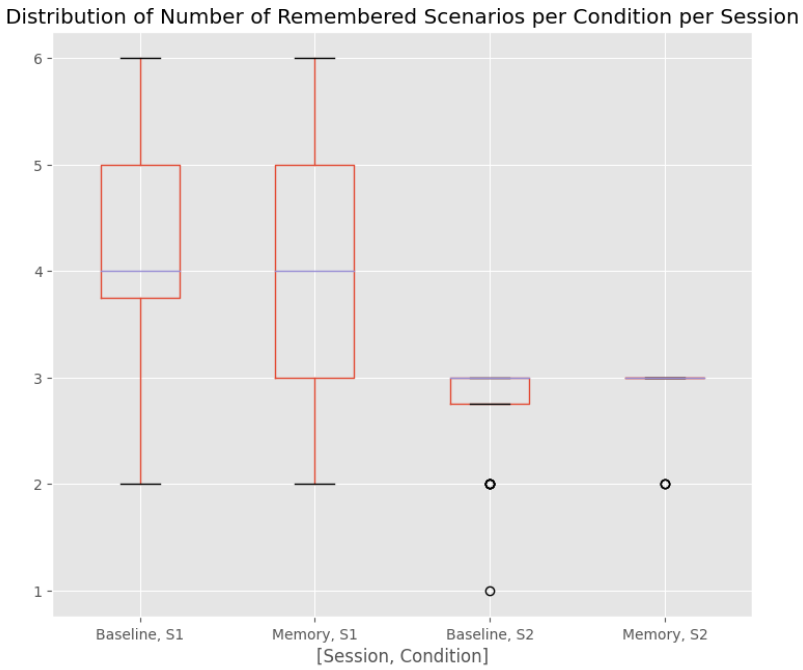

Figure 5.15: Distribution of Number of Remembered Scenarios per Session per Condition

Looking at the Quality Level of those scenarios in Figure 5.16, along with the percentage of those that were not mentioned, participants seem to remember the scenarios more accurately and for the most part can describe the images. In session 1, participants in the baseline condition seem to recollect scenarios with high quality more than those in the memory condition (55.62% vs 48.15%), while in session 2 the opposite is observed (82.56% vs 86.59%).

To better study the quality of recollections, the average quality per participant is calculated according to the enumeration explained in section 4.5. In Figure 5.17, the distribution of the participants' average quality level is shown. The Shapiro-Wilk normality test reveals

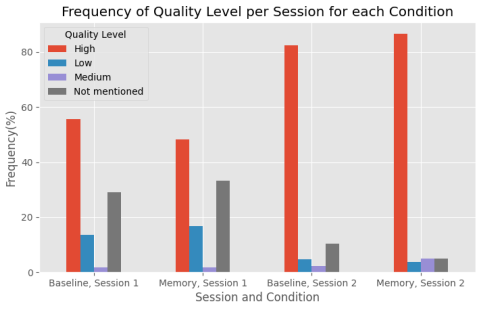

Figure 5.16: Frequency of Quality Levels per Session per Condition

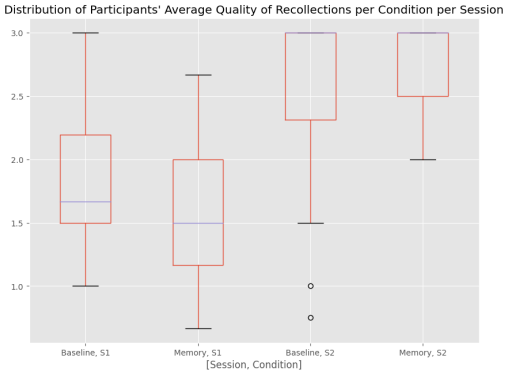

Figure 5.17: Distribution of Participants' Average Quality of Recollections per Session per Condition

that the data does not follow a normal distribution ($W = 0.66, p << 0.001$). Performing a Kruskal-Wallis test, we conclude that the quality average is not significantly affected by the condition ($H(1) = 0.047, p = 0.828$).

At this point, it is also interesting to look at the reasons that participants provided for remembering specific scenarios. The frequency of all the sub-labels per condition and the categories of labels per condition and session can be found in Figures 5.18 and 5.19 respectively.

It is worth mentioning that the 'Feelings' category is much more prominent during session 1, while 'Thought & Ability' rises in session 2. That makes sense considering that the questions in session 2 are more complicated and require participants to reflect more within themselves. A chi-square test is attempted to explore the relation between the quality of recollections and the reason for remembering, but due to the low number of 'medium' level recollections, it is required to merge that level with another, which will be the 'low' level, since those recollections still fail to provide an accurate description of the pictures. The test shows a significant association between the categories of recollections and the quality ($\chi^2(4) = 11.89, p = 0.0181$). The full contingency table can be found in Figure 5.20, as well as the respective mosaic plot (Figure 5.21). From the mosaic plot, it becomes apparent that participants who were not able to describe the images accurately or at all, provided a reason

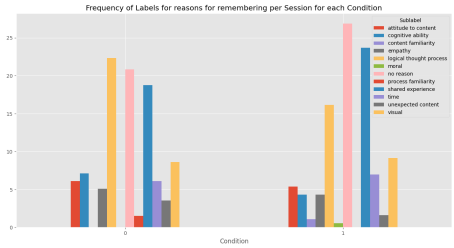

Figure 5.18: Frequency of Sub Labels per Condition

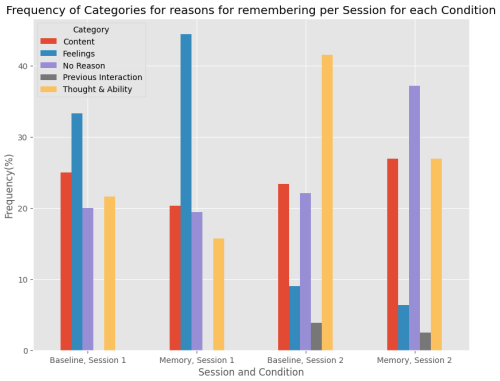

Figure 5.19: Frequency of Categories of Labels per Condition

from the category 'Feelings' as the main reason for remembering scenarios, while those that could detail the images, considered 'Content' the major factor. Comparing the two sessions, it appears that 'Session' has a big impact on reasons for remembering, not allowing room for 'Condition' to have an effect.

	mem$Quality_Num		
mem$Category	1	2	Row Total
Content	6	85	91
	16.394	74.606	
	6.593%	93.407%	23.760%
	8.696%	27.070%	
	-2.567	1.203	
Feelings	25	75	100
	18.016	81.984	
	25.000%	75.000%	26.110%
	36.232%	23.885%	
	1.646	-0.771	
No Reason	19	72	91
	16.394	74.606	
	20.879%	79.121%	23.760%
	27.536%	22.930%	
	0.644	-0.302	
Previous Interaction	1	4	5
	0.901	4.099	
	20.000%	80.000%	1.305%
	1.449%	1.274%	
	0.105	-0.049	
Thought & Ability	18	78	96
	17.295	78.705	
	18.750%	81.250%	25.065%
	26.087%	24.841%	
	0.170	-0.079	
Column Total	69	314	383
	18.016%	81.984%	

Figure 5.20: Contingency Table for chi-square test between 'Category' of recollection and Quality

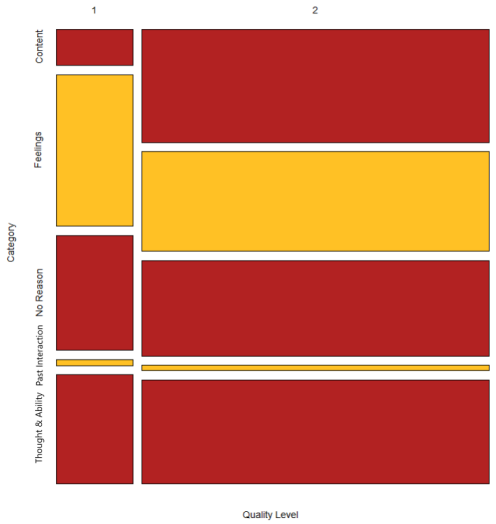

Figure 5.21: Mosaic of the categories of recollections per quality levels

Chapter 6

Discussion

6.1 Implications

After the analysis of the data, the findings and implications of this study can now be discussed regarding the impact of engagement strategies and memory on understanding during human-robot interactions in education. No previous study was found that directly correlated the quality of argumentation and engagement strategies (*H1*, 2.4). Our findings suggest that initially there is a significant effect between the depth of participants' arguments and the use of engagement strategies. The hypothesis can be confirmed regarding the first interaction but not in its entirety since the effect disappears in the second interaction, and additionally, no effect was noticed on the Breadth level.

No relation was detected between the employment of engagement strategies and the participants' ability to correctly identify the value profile extracted from the robot's memory model (*H2*), with more people choosing the incorrect one. A very similar study was implemented by Saveur[63], using a different memory model and taking into account all of Schwartz's values. In their study, almost all participants interacting with the complete memory condition were able to identify the robot's plot as the closest to their own. This leads to the exploration of whether the value profile was accurate and if participants agreed with the robot in its assumptions (*H3*). There is a slight indication that for the participants who chose the plot generated by the robot's memory model, the value assumptions made by the robot and the values displayed by their arguments tended to coincide more often.

Nevertheless, it is not enough to confirm the hypothesis and a more complete value model like the one employed by Saveur might be necessary.

Likability in human-robot interaction has conflicting results in the existing literature. Some studies have been able to show an increase in the likability of the robot by using a memory model and incorporating shared experiences [34, 37]. Leite, Pereira, and Lehman[36] noticed mixed results, depending on the age of the participants, where younger children preferred the control version, and older the enhanced one. Saravanan et al.[31] observed no notable change to any of the Godspeed sections. Our hypothesis about the connection between personalization in the memory condition and likeability (*H4*) cannot be confirmed by the data, and no significant effect can be reported for any of the other sectors as well. Given the very high ratings of likability, the novelty effect could be an explanation for the lack of difference between the two conditions.

The last hypothesis regarding the recollection quality being increased for the memory condition (*H5*) cannot be confirmed based on the analysis of the data. When also considering the reasons for remembering, there is a clear pattern of different reasons being more prominent for each session, indicating that the effect of the session was too strong to allow room for the condition to have any effect. Similar studies that measure information retention during the interaction mostly focus on easily measurable outcomes, such as words learned[38] or mathematics score[64], rely on quantity over quality and have been shown to increase the learning results. Quality, however, is not so easy to quantify, and although there are studies that explore the quality of recollections[46, 65], it is not in regards to engagement strategies, therefore it is not possible to compare them.

6.2 Limitations

6.2.1 Children-centered experiment with Adults

The study was initially designed for children, as the pilot was also performed with younger ages. However, there were challenges in securing participation, either because the parents were hesitant to give consent, the school's scheduling did not allow for it, or because of the

language barrier; the study was conducted in English in the Netherlands, so an international or bilingual school was required. The only successful attempt at procuring children participants led to 6 children with informed consent, of which only 3 completed the experiment in full, due to technical issues, scheduling conflicts, or issues with understanding the language. Thus, the sample was deemed too small to deduce any results of statistical significance, leading to the decision to conduct the experiment with adults.

Since the topic and visual content were tailored to a child audience, and despite the instruction that was given to answer the questions from the position of their current self, and not of their past, child self, it is acknowledged that adults may not find the materials as relatable. Additionally, the inherent differences in behavior, mental capacity, and communication style between adults and children raise concerns about the generalizability of the findings to the intended target audience. Moreover, adults may articulate their thoughts differently, potentially influencing the quality of the collected data.

6.2.2 Non-diverse Population

The participant population is another limitation, primarily consisting of university students and employees, resulting in a sample that is highly educated and potentially not representative of a more diverse demographic. The study's location at a technical university contributed to a skewed gender distribution, with a larger male population. This may limit the broader applicability of the study's results, particularly in contexts with different educational and gender compositions. Furthermore, while only 14 of the 55 participants had previous experience with the NAO robot, the technological background of many of the participants implies a certain familiarity with newer technologies, that might not be present in a different demographic.

6.2.3 Recollections Quality

The labeling scale for the quality of recollections also poses limitations. Due to the impracticality and the time investment of individually annotating interviews with the proposed scheme by Nikkels[46], a different scale was devised. However, this alternative system may

lack the necessary granularity to effectively differentiate between types of recollections and provide any statistical significance.

6.2.4 Time of Interaction

Finally, the participants' availability introduced a constraint on the experiment's scheduling, with sessions taking place at varying hours of the day. This flexibility, while practical, may have influenced the participants' attention and receptivity during the sessions, considering the demonstrated variation in learning effectiveness at different times of the day, impacting the overall quality of the collected data.

6.3 Further Research

6.3.1 Value Profile

One avenue for further research could involve exploring improved methods for creating the value profiles and assessing and comparing values within the context of the study. Currently, the restricted number of values potentially limit the accuracy of representation in the Bayesian network. Furthermore, another form of semantic memory could be considered, since the updates to the Bayesian Network don't seem to suffice to create a reflection of the participants' values. Expanding the range of values to capture a more comprehensive spectrum, would enable a more nuanced understanding of participants' value profiles. Additionally, investigating ways to incorporate the importance of context into the information conveyed to the participants could enhance their understanding of their values.

6.3.2 Better Recollection Quality Scheme

Enhancing the labeling scheme for recollections represents another area of potential research. The current challenges in capturing levels of remembering, particularly in a large sample size, suggest a need for a more practical and efficient system. Future research could explore the development of a labeling system that allows for better granularity in assessing recollection quality, while also being feasible for large-scale experiments. An automated or

semi-automated system that can be filled during the interview process might offer a more realistic and scalable approach to data annotation.

6.3.3 Visual Expressions

Given the participants' feedback on the impact of expressions in images on their understanding and decision-making, there is potential for further research into visual messages and their role in information retention. Exploring how facial expressions influence participants' perception of scenarios and contribute to memory retention could offer valuable insights. This line of research could involve controlling facial expressions in images to observe their effects on decision-making and recollection. Understanding the interaction between visual cues and cognitive processes may contribute to the development of more effective visual communication strategies within the context of interactive e-health solutions.

Chapter 7

Conclusion

In this study, the main research theme was how a robot enhanced with engagement strategies affects the understanding and the retention capacity for information when reflecting on personal values. The robot discussed with the participants their behavior choices in different situations and attempted to improve their understanding of the topic in general and of their values.

In the context of this research, the effect of the engagement strategies between sessions was studied in terms of the quality of the arguments the participants provided. A unique scheme was developed for this purpose. Additionally, the accuracy of the robot's assumptions regarding the behavior choices is tested through plots reflecting the importance of the values and a combination of self-reported agreements and annotated-assumed value consensus. The behavioral impact of the robot is also studied, and more specifically its likability. Finally, the effect of the engagement strategies on information retention is explored and the quality and reasons for remembering are analyzed and linked.

Even though the experiment was not possible to be performed with children at this time, several takeaways can be utilized in similar educational contexts. A significant improvement between the two conditions was not found but some trends were identified that provide insight for further research. Participants in the memory condition seem to be more confident regarding their Plot Choice which could indicate a deeper understanding of their values. They also seem to not find the robot as likable, but they do perceive it to be more intelligent. That

could be attributed to the sharing of previous experiences, where the robot will reference the participants' arguments in the next session. The quality of recollections was much higher for the second session since only 3 scenarios were discussed, compared to the 6 of the first session, explained by the less cognitive stress, and in some cases the prior knowledge of the process. An interesting connection was made between two of the categories of reasons for remembering and the quality of recollections: in cases where the subject was not able to recall the image displayed on the screen, the reason for remembering was mostly attributed to 'Feeling', while when they provided an accurate description of the picture, the most probable reason was 'Content'.

In conclusion, this study contributes to the growing body of research investigating the influence of engagement strategies employed by robots in educational contexts. It not only introduces an effective scheme for assessing argument quality as a learning outcome but also offers meaningful insights into the nuances of attention and engagement dynamics during human-robot interactions. By examining the reasons for retaining information and their implications for conversation points, this study provides valuable guidance on how to strategically capture attention and maintain engagement in human-robot interactions. These findings offer a better understanding of the relationship between engagement and learning outcomes, paving the way for more refined approaches in designing effective and engaging human-robot educational interactions.

Bibliography

[1] Hansol Woo et al. "The use of social robots in classrooms: A review of field-based studies". In: *Educational Research Review* 33 (2021), p. 100388. ISSN: 1747-938X. DOI: https://doi.org/10.1016/j.edurev.2021.100388. URL: https://www.sciencedirect.com/science/article/pii/S1747938X21000117.

[2] Tony Belpaeme et al. "Social robots for education: A review". In: *Science Robotics* 3.21 (2018), eaat5954. DOI: 10.1126/scirobotics.aat5954. eprint: https://www.science.org/doi/pdf/10.1126/scirobotics.aat5954. URL: https://www.science.org/doi/abs/10.1126/scirobotics.aat5954.

[3] Rianne van den Berghe et al. "Social Robots for Language Learning: A Review". In: *Review of Educational Research* 89.2 (2019), pp. 259–295. DOI: 10.3102/0034654318821286. URL: https://doi.org/10.3102/0034654318821286.

[4] George A. Papakostas et al. "Social Robots in Special Education: A Systematic Review". In: *Electronics* 10.12 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10121398. URL: https://www.mdpi.com/2079-9292/10/12/1398.

[5] Candace L. Sidner et al. "Explorations in engagement for humans and robots". In: *Artificial Intelligence* 166.1 (2005), pp. 140–164. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2005.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0004370205000512.

[6] Catharine Oertel et al. "Engagement in Human-Agent Interaction: An Overview". In: *Frontiers in Robotics and AI* 7 (2020). ISSN: 2296-9144. DOI: 10.3389/frobt.2020.00092. URL: https://www.frontiersin.org/articles/10.3389/frobt.2020.00092.

[7] Rebecca Strachan and Lalith Liyanage. "Active Student Engagement: The Heart of Effective Learning". In: *Global Innovation of Teaching and Learning in Higher Education: Transgressing Boundaries*. Ed. by Prudence C. Layne and Peter Lake. Cham: Springer International Publishing, 2015, pp. 255–274.

[8] Chris Lytridis et al. "On Measuring Engagement Level During Child-Robot Interaction in Education". In: Jan. 2020, pp. 3–13. ISBN: 978-3-030-26944-9. DOI: 10.1007/978-3-030-26945-6_1.

[9] Jauwairia Nasir et al. "What if Social Robots Look for Productive Engagement?: Automated Assessment of Goal-Centric Engagement in Learning Applications". In: *International Journal of Social Robotics* 14 (Jan. 2022). DOI: 10.1007/s12369-021-00766-w.

[10] Shalom H. Schwartz. "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries". In: ed. by Mark P. Zanna. Vol. 25. Advances in Experimental Social Psychology. Academic Press, 1992, pp. 1–65. DOI: https://doi.org/10.1016/S0065-2601(08)60281-6. URL: https://www.sciencedirect.com/science/article/pii/S0065260108602816.

[11] Thomas Hughes-Roberts et al. "Examining engagement and achievement in learners with individual needs through robotic-based teaching sessions". In: *British Journal of Educational Technology* 50.5 (2019), pp. 2736–2750. DOI: https://doi.org/10.1111/bjet.12722. eprint: https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12722. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12722.

[12] Mirjam de Haas, Paul Vogt, and Emiel Krahmer. "The Effects of Feedback on Children's Engagement and Learning Outcomes in Robot-Assisted Second Language Learning". In: *Frontiers in Robotics and AI* 7 (2020). ISSN: 2296-9144. DOI: 10.3389/frobt.2020.00101. URL: https://www.frontiersin.org/articles/10.3389/frobt.2020.00101.

[13] Shalom H. Schwartz. "Are There Universal Aspects in the Structure and Contents of Human Values". In: *Journal of Social Issues* 50 (1994), pp. 19–45. URL: https://api.semanticscholar.org/CorpusID:15121950.

[14] Rebecca Collie and Andrew Martin. "Motivation and Engagement in Learning". In: Dec. 2019. DOI: 10.1093/acrefore/9780190264093.013.891.

[15] Richard Ryan and Edward Deci. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being". In: *The American psychologist* 55 (Feb. 2000), pp. 68–78. DOI: 10.1037/0003-066X.55.1.68.

[16] W.R. Miller and S. Rollnick. *Motivational Interviewing: Helping People Change*. Applications of Motivational Interviewing Series. Guilford Publications, 2012. ISBN: 9781609182274. URL: https://books.google.nl/books?id=o1-ZpM7QqVQC.

[17] Daniel Schulman, Timothy Bickmore, and Candace Sidner. "An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change Using Motivational Interviewing." In: Jan. 2011.

[18] Toshikazu Kanaoka and Bilge Mutlu. "Designing a Motivational Agent for Behavior Change in Physical Activity". In: *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '15. , Seoul, Republic of Korea, Association for Computing Machinery, 2015, pp. 1445–1450. ISBN: 9781450331463. DOI: 10.1145/2702613.2732924. URL: https://doi.org/10.1145/2702613.2732924.

[19] Samiha Samrose and Ehsan Hoque. "MIA: Motivational Interviewing Agent for Improving Conversational Skills in Remote Group Discussions". In: *Proc. ACM Hum.-Comput. Interact.* 6.GROUP (Jan. 2022). DOI: 10.1145/3492864. URL: https://doi.org/10.1145/3492864.

[20] Martyn Standage, Joan Duda, and Nikos Ntoumanis. "Students' Motivational Processes and Their Relationship to Teacher Ratings in School Physical Education: A Self-Determination Theory Approach". In: *Research quarterly for exercise and sport* 77 (Mar. 2006), pp. 100–10. DOI: 10.1080/02701367.2006.10599336.

[21] Irwin Altman and Dalmas A. Taylor. *Social penetration: The development of interpersonal relationships*. Holt, Reinhart and Winston, 1973.

[22] Amanda Carpenter and Kathryn Greene. "Social Penetration Theory". In: Dec. 2015. DOI: 10.1002/9781118540190.wbeic160.

[23] Franziska Burger, Joost Broekens, and Mark A. Neerincx. "Fostering Relatedness Between Children and Virtual Agents Through Reciprocal Self-disclosure". In: *BNAIC 2016: Artificial Intelligence*. Ed. by Tibor Bosse and Bert Bredeweg. Cham: Springer International Publishing, 2017, pp. 137–154. ISBN: 978-3-319-67468-1.

[24] Peggy van Minkelen et al. "Using Self-Determination Theory in Social Robots to Increase Motivation in L2 Word Learning". In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, 2020, pp. 369–377. ISBN: 9781450367462. DOI: 10.1145/3319502.3374828. URL: https://doi.org/10.1145/3319502.3374828.

[25] Richard Koestner, Miron Zuckerman, and Julia Koestner. "Praise, involvement, and intrinsic motivation." In: *Journal of Personality and Social Psychology* 53 (1987), pp. 383–390.

[26] Shannon Zentall and Bradley Morris. "" Good job, you're so smart": The effects of inconsistency of praise type on young children's motivation". In: *Journal of experimental child psychology* 107 (Oct. 2010), pp. 155–63. DOI: 10.1016/j.jecp.2010.04.015.

[27] Bradley Morris and Shannon Zentall. "High fives motivate: The effects of gestural and ambiguous verbal praise on motivation". In: *Frontiers in Psychology* 5 (Aug. 2014), pp. 1–6. DOI: 10.3389/fpsyg.2014.00928.

[28] Mike E.U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. "Memory-Based Personalization for Fostering a Long-Term Child-Robot Relationship". In: vol. 2022-March. 2022, pp. 80–89. DOI: 10.1109/HRI53351.2022.9889446.

[29] Raquel Ros et al. "Child-robot interaction in the wild: advice to the aspiring experimenter". In: *Proceedings of the 13th International Conference on Multimodal Interfaces*. ICMI '11. Alicante, Spain: Association for Computing Machinery, 2011, pp. 335–342. ISBN: 9781450306416. DOI: 10.1145/2070481.2070545. URL: https://doi.org/10.1145/2070481.2070545.

[30] Iolanda Leite, Carlos Martinho, and Ana Paiva. "Social Robots for Long-Term Interaction: A Survey". In: *International Journal of Social Robotics* 5 (2013), pp. 291–308. URL: https://api.semanticscholar.org/CorpusID:3721600.

[31] Avinash Saravanan et al. "Giving Social Robots a Conversational Memory for Motivational Experience Sharing". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2022, pp. 985–992. DOI: 10.1109/RO-MAN53752.2022.9900677.

[32] Mike E.U. Ligthart et al. "Back to School - Sustaining Recurring Child-Robot Educational Interactions After a Long Break". In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 433–442. ISBN: 9798400703225. DOI: 10.1145/3610977.3635001. URL: https://doi.org/10.1145/3610977.3635001.

[33] Ivana Kruijff-Korbayova et al. "Young Users' Perception of a Social Robot Displaying Familiarity and Eliciting Disclosure". In: Oct. 2015, pp. 380–389. ISBN: 978-3-319-25553-8. DOI: 10.1007/978-3-319-25554-5_38.

[34] Changzeng Fu et al. "Sharing Experiences to Help a Robot Present Its Mind and Sociability". In: *International Journal of Social Robotics* 13 (Apr. 2021). DOI: 10.1007/s12369-020-00643-y.

[35] Changzeng Fu et al. "Using an Android Robot to Improve Social Connectedness by Sharing Recent Experiences of Group Members in Human-Robot Conversations". In: *IEEE Robotics and Automation Letters* PP (July 2021), pp. 1–1. DOI: 10.1109/LRA.2021.3094779.

[36] Iolanda Leite, André Pereira, and Jill Fain Lehman. "Persistent Memory in Repeated Child-Robot Conversations". In: *Proceedings of the 2017 Conference on Interaction Design and Children*. IDC '17. Stanford, California, USA: Association for Computing Machinery, 2017, pp. 238–247. ISBN: 9781450349215. DOI: 10.1145/3078072.3079728. URL: https://doi.org/10.1145/3078072.3079728.

[37] Nikhil Churamani et al. "The Impact of Personalisation on Human-Robot Interaction in Learning Scenarios". In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI '17. Bielefeld, Germany: Association for Computing Machinery, 2017, pp. 171–180. ISBN: 9781450351133. DOI: 10.1145/3125739.3125756. URL: https://doi.org/10.1145/3125739.3125756.

[38] Muneeb Ahmad, Omar Mubin, and Joanne Orlando. "Adaptive Social Robot for Sustaining Social Engagement during Long-Term Children–Robot Interaction". In: *International Journal of Human-Computer Interaction* 33 (Mar. 2017), pp. 1–20. DOI: 10.1080/10447318.2017.1300750.

[39] Muneeb Ahmad et al. "Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: a long-term child–robot interaction study". In: *Adaptive Behavior* 27 (May 2019), p. 105971231984418. DOI: 10.1177/1059712319844182.

[40] Endel Tulving. *Elements of Episodic Memory*. Oxford University Press, 1983.

[41] Endel Tulving et al. "Episodic and semantic memory". In: *Organization of memory* 1.381-403 (1972), p. 1.

[42] Wan Ho et al. "An Initial Memory Model for Virtual and Robot Companions Supporting Migration and Long-term Interaction". In: Nov. 2009, pp. 277–284. DOI: 10.1109/ROMAN.2009.5326204.

[43] Zerrin Yumak and Nadia Thalmann. "Towards episodic memory-based long-term affective interaction with a human-like robot". In: Oct. 2010, pp. 452–457. DOI: 10.1109/ROMAN.2010.5598644.

[44] Taewoon Kim et al. "A Machine with Short-Term, Episodic, and Semantic Memory Systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.1 (June 2023), pp. 48–56. DOI: 10.1609/aaai.v37i1.25075. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25075.

[45] John Rauthmann et al. "The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics". In: *Journal of Personality and Social Psychology* 107 (Sept. 2014), pp. 677–718. DOI: 10.1037/a0037250.

[46] Lucile Nikkels. *Memorable moment detection using eye gaze in child-robot interactions*. Master's thesis. Available at http://resolver.tudelft.nl/uuid:90fee300-8978-4b3b-9ece-3af6b1ea6dc4. Delft, The Netherlands, Aug. 2023.

[47] Vladimir Ponizovskiy et al. "Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text". In: *European Journal of Personality* 34.5 (2020), pp. 885–902. DOI: https://doi.org/10.1002/per.2294. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/per.2294. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/per.2294.

[48] Hossein Banaeian and Ilkay Gilanlioglu. "Influence of the NAO robot as a teaching assistant on university students' vocabulary learning and attitudes". In: *Australasian Journal of Educational Technology* 37.3 (Apr. 2021), pp. 71–87. DOI: 10.14742/ajet.6130. URL: https://ajet.org.au/index.php/AJET/article/view/6130.

[49] Helen Crompton, Kristen Gregory, and Diane Burke. "Humanoid robots supporting children's learning in an early childhood setting". In: *British Journal of Educational Technology* 49.5 (2018), pp. 911–927. DOI: https://doi.org/10.1111/bjet.12654. eprint: https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12654. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12654.

[50] Joana Galvão Gomes da Silva et al. "Experiences of a Motivational Interview Delivered by a Robot: Qualitative Study". In: *J Med Internet Res* 20.5 (May 2018), e116. ISSN: 1438-8871. DOI: 10.2196/jmir.7737. URL: https://doi.org/10.2196/jmir.7737.

[51] Stefan Heinrich and Stefan Wermter. "Towards robust speech recognition for human-robot interaction". In: *Proceedings of the IROS2011 Workshop on Cognitive Neuroscience Robotics (CNR)*. 2011, pp. 29–34.

[52] Barry Zimmerman. "Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects". In: *American Educational Research Journal - AMER EDUC RES J* 45 (Mar. 2008), pp. 166–183. DOI: 10.3102/0002831207312909.

[53] Ernesto Panadero. "A Review of Self-regulated Learning: Six Models and Four Directions for Research". In: *Frontiers in Psychology* 8 (2017). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.00422. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00422.

[54] B.J. Zimmerman and Adam Moylan. "Self-regulation: Where metacognition and motivation intersect". In: *Handbook of metacognition in education* (Jan. 2009), pp. 299–315.

[55] J.B. Biggs, K.F. Collis, and A.J. Edward. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. Elsevier Science, 2014. ISBN: 9781483273310. URL: https://books.google.nl/books?id=xUO0BQAAQBAJ.

[56] Thomas Daniel Ullmann. "Automated detection of reflection in texts: a machine learning based approach". In: 2015. URL: https://api.semanticscholar.org/CorpusID:61766690.

[57] Ellen Isaacs et al. "Echoes From the Past: How Technology Mediated Reflection Improves Well-Being". In: Apr. 2013. DOI: 10.1145/2470654.2466137.

[58] Mark Snyder. "Self-monitoring of expressive behavior." In: *Journal of Personality and Social Psychology* 30 (1974), pp. 526–537. URL: https://api.semanticscholar.org/CorpusID:144979719.

[59] Richard D. Lennox and Raymond N. Wolfe. "Revision of the self-monitoring scale." In: *Journal of personality and social psychology* 46 6 (1984), pp. 1349–64. URL: https://api.semanticscholar.org/CorpusID:8668419.

[60] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International Journal of Social Robotics* 1.1 (2009), pp. 71–81. DOI: 10.1007/s12369-008-0001-3.

[61] H. Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, 1885. URL: https://books.google.nl/books?id=kfAOAAAAMAAJ.

[62] Maria Tsfasman et al. "Towards creating a conversational memory for long-term meeting support: predicting memorable moments in multi-party conversations through eye-gaze". In: Nov. 2022, pp. 94–104. DOI: 10.1145/3536221.3556613.

[63] Tom Saveur. *Contextualised Value Model*. Master's thesis. Available at http://resolver.tudelft.nl/uuid:566ff230-4eae-48b9-8ecc-83a7e981babe. Delft, The Netherlands, Mar. 2024.

[64] Muneeb Ahmad and Omar Mubin. "Emotion and Memory Model to Promote Mathematics Learning - An Exploratory Long-term Study". In: Dec. 2018, pp. 214–221. DOI: 10.1145/3284432.3284451.

[65] Laura Ottevanger. *Exploring Children's Choices in an Educational Game on Neurodiversity: Reveiling Underlying Values through Robot's Socratic Questioning*. Master's thesis. Available at http://resolver.tudelft.nl/uuid:a7e8d023-f990-4bfc-afa6-a8cfbb536c30. Delft, The Netherlands, Sept. 2023.

Appendix A

Scenarios

Value Conflict	**D.I.A.**	**Scenario description**
Self-Direction vs Benevolence	0,0,1	you play tag and your friend denies that you tapped her
Self-Direction vs Benevolence	0,0,1	you are falsely accused of stealing a pen
Achievement vs Benevolence	0,1,0	you and your best friend have to fill in a worksheet but they don't understand it while you find it easy
Achievement vs Benevolence	0,1,0	the person sitting next to you can't concentrate well during a math course and starts distracting you
Conformity vs Self-Direction	0,1,1	you have to prepare a group presentation but you disagree with the topic chosen by the group
Conformity vs Self-Direction	0,1,1	you think the professor is wrong
Benevolence vs Conformity	1,0,0	you see a classmate being bullied but your friends seem to find it funny
Benevolence vs Conformity	1,0,0	a friend asks you if he can copy your assignment
Self-Direction vs Achievement	1,1,0	you have trouble concentrating on your test and an opportunity presents itself to peer at the person sitting next to you
Self-Direction vs Achievement	1,1,0	you saw someone getting hurt but had to run back to class for an exam
Conformity vs Achievement	1,1,1	you know a lot about the teaching material and the professor asks you to help her with the teaching
Conformity vs Achievement	1,1,1	your friends would like to go to an environment strike but the professor also wanted to teach an important lesson

Table A.1: Full list of scenarios used

Appendix B

Interview Script

B.1 Free recall Section

I will now ask you a few questions about the conversation you just had with Robin.

1. Can you tell me some things that you remember from the discussion with the robot? *If the answer is not specific enough, rephrase the question, e.g.*
   - Do you remember any of the situations that you discussed?
2. Do you remember a specific moment when this happened?
   - Can you tell me some more details from this moment?
   - Do you remember the scenario/decision that you were discussing at that time?
3. Do you remember what was on the screen during that particular moment?

   If not enough details are given, prompt for more information
   - Can you describe the background/people/actions?
4. Why do you think that you remember this moment in particular?
   - Do you think there is a specific reason that you remember this one better?

   Try to identify whether the 'why' is related to:
   - *the robot's behavior*

- *the visuals on screen*
- *the content/decision making process*

5. Was there anything else that caught your attention and you would like to mention? *This can be about the interaction, the robot, the pictures, or even their thinking process. An open question for free expression of thought.*

B.2 Visual Stimuli Section

1. *Pull out copies of the scenario visuals that they were shown*
   - Can you tell me which of these images you remember the best and why?
2. Is there anything else that you noticed? *Another moment for free expression.*