Memorable moment detection using eye gaze in child-robot interactions

L.R.M. Nikkels

Memorable moment detection using eye gaze in child-robot interactions

by

L.R.M. Nikkels

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday August 30, 2023 at 01:00 PM.

Student number:	4564804	
Project duration:	Nov 14, 2022 – Aug 30, 2023	
Thesis committee:	Dr. C. R. M. M. Oertel,	TU Delft, supervisor
	Prof. Dr. M. A. Neerincx,	TU Delft, supervisor
	Dr. X. Zhang,	TU Delft

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

TU Delft

Abstract

Robots in classroom settings can help teachers with providing personalised attention to children's health and development. As part of this personalisation, robots should store and use (verbal or multi-modal) information about the children they interact with. One aspect that has been unexplored in existing literature is the detection of memorable moments during these child-robot interaction. Eye-gaze tracking is a low cost and non-invasive method applied widely to gain insight into human's inner processes. This study has found that several state-of-the-art time series machine learning models perform better-than-chance on the detection of memorable moments using gaze tracking. In addition, a shapelet-based transform classifier also performed better-than-chance in distinguishing memories according to 3 different levels of recall detail. Manual data analysis has identified significantly different gaze behaviour during memorable moments and not memorable moments as well as in the gaze behaviour for different levels of recall detail. The comparison of the results with related literature leads to the hypothesis that memorable moments are likely to be moments of both high levels of engagement and deep thinking. The data analyses also provided insight into children's gaze behaviour for different reasons for remembering a moment. The results show that these reasons, or 'internal processes', are distinguishable by gaze patterns and thus provide insight into items or concept that draw the child's attention. This study shows that memorable moments detection for children is a developing and promising field that could potentially provide a lot of insight into children's situated thought processes.

Preface

This document contains the thesis that is completed as part of my graduation project with the Interactive Intelligence group at TU Delft. At the start of this project I had no idea about the magnitude and intensity of such a project. I threw myself in the deep end and was happy to learn as much as I did about different AI techniques and their relation to children, but mostly about how to conduct a professional research. Now I know about the complexities of setting up experiments, the perseverance it takes to dig through endless papers and the satisfaction of feeling like I have found something meaningful. I hope that with all of my efforts, I have made a contribution to the field of research that will help to enhance society by enabling children with more attention and care.

This thesis would not have come to exist without the continuous help and involvement of several people. In particular, I would like to thank and acknowledge my supervisors Catharine Oertel and Mark Neerincx. Without their weekly feedback and support I would not have found the direction and depth in my research that I managed to obtain. I would like to give special thanks to Catharine Oertel for being as involved and caring as she was, during all circumstances, which made me feel both well-guided and taken care of. I want to thank Xucong Zhang for agreeing to be in my thesis committee. In addition, I would like to thank Masha Tsfasman for being a support both content-wise and emotionally. For enabling and setting up the pillars of my experiment, I want to thank especially Fran Burger but also all other people that were involved in the process. My research would not have been made possible without the help of my fellow students as well. A special thanks goes out to Laura Ottenvanger for helping during my experiment, Marciano Jorden for helping me process my data and Tom Saveur for reviewing my data labelling. I would also like to acknowledge the ePartners4All project and its partners 4TU for providing the support and funding that allowed for this research to be completed. Finally, I want to thank the people close to me, both in the Netherlands and in Berlin, for taking care of me and supporting me during this whole process.

L.R.M. Nikkels
Delft, August 2023

Contents

1 Introduction

1.1. Context

Over the last 20 years, technology has made great advancements and has now become inevitably intertwined with human lives. Modern technology can be used in all areas of life to ease or enhance task execution. Particularly, in the field of education, technology such as smartboards, laptops and tablets have provided easy, fun and engaging opportunities for new ways of teaching and learning (Harper & Milman, 2016). A challenge that remains present in the education system is improving children's mental and physical health (Reinke et al., 2011). Especially children from lower socio-economic backgrounds or children with special needs could gain a lot, health-wise, from getting more attention and/or individual attention (Barragán-Sánchez et al., 2023; Cappella et al., 2008). Human educators are very capable to provide this kind of support, but are falling short in terms of available time, energy and money to invest in this (Reinke et al., 2011). This provides an opportunity for modern technology to come in and take some of that effort out of their hands, by improving individual attention for children at a relatively low cost (Hasselbring & Glaser, 2000). This opportunity has been previously identified by numerous different (research) institutions in the world. A group of European institutions came together in 2021 to start a project in this field of research, called 'ePartners4All' ("ePartners4All", 2021).

ePartners4All

On their website[1] the ePartners4All project is described as follows:

> "In this project we take digital support of school-aged children and their caregivers a big leap forward, by not only monitoring their health, but also by providing privacy proof and mutual accepted and co-developed interactive e-health solutions (so-called ePartners), including robot buddies and virtual agents that enhance children's health and well-being. There is both a large knowledge base in each individual partner, and much more to learn from each other and by combining our knowledge and technologies. These ePartners can help to prevent health problems in at-risk children, and it can help to recognize and treat health problems at an early-stage, thereby preventing deterioration of the problem. In this way, ePartners4All can help to create a more resilient

Figure 1.1: The NAO robot, as used in the ePartners4All project and this study.

[1] https://epartners4all.com/ePartners4All

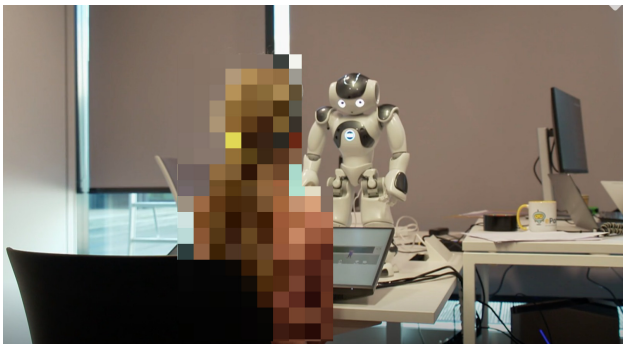


Figure 1.2: The experiment setup, including (anonymized) participant, robot, and screen with visual aid.

Figure 1.3: An example visual aid image. In this example the main character (girl in blue) tries to convince the boy with the pancakes to swap their lunch.

> society. Altogether, it could lead to lower healthcare utilization and, in the long run, a more resilient workforce with lower losses of productivity."

The project has different components and is being carried out by different, international parties such as research organisations, universities and private companies. The role of TU Delft in this project is to develop the functional component, i.e. conversational memory, of a robot buddy and to evaluate this component with school children. The robot that is being used for this experiment is the NAO robot, depicted in Figure 1.1.

During the child-robot interaction directed by TU Delft, the robot aims to learn about the child's values, such as "kindness" and "performance" (Schwartz, 2006). The information about the children's values can be used by intelligent systems to get a deeper understanding of the children. With this information, a system could identify in what areas children might need extra guidance or attention. In addition, it can also be used as a tool to relate to children and motivate them in effective ways, using these values. Finally, a discussion to reflect on decisions and their consequences can teach children to be more aware of their own and others' inner monologue. A system that can infer the children's visual attention and relate this to their chance of remembering the moment could potentially infer these preferences before they mention it themselves or explicitly. This could facilitate easier and more enjoyable ways to obtain this information. In addition, it could serve as an extra verification of the findings. During the experiment, the children are asked to make decisions in fictional everyday (school-related) situations that are guided by these values (see Figure 1.2 and Figure 1.3). In a subsequent session, the child and robot continue the conversation and reflect on or revisit the earlier choices. For this, a memory is needed, but current robots lack a memory that can drive such a conversation (Campos et al., 2018).

TU Delft has developed a conversational memory for the NAO robot, such that it is granted the ability to remember these relevant moments and decisions. The value-choice experiment will be used to test the validity and the usefulness of this episodic conversational memory implementation. The children will reflect on their value choices from different points of view throughout multiple sessions, which is made possible with the implementation of this conversational memory. During these reflection sessions, the robot could relate or refer to the children's personal experience based on previous information, and hereby make them feel heard and understood. Especially if children lack attention in their home environment, this could contribute positively to their mental health (Abbasi et al., 2022).

On the long term, through learning about children's inner values, the robot aims to improve the children's value awareness. This means that children would be better able to identify and explain ethical or moral values and to relate this to their own or other people's behavioral choices. A desirable outcome of this process is that the children will be better able to understand different points of view and handle socially challenging situations. This (applied) knowledge benefits the children's health in the way that it can relieve stress and tension within social relationships both in the classroom and at home. Finally, the gathered data in children's preferences and behavioral decisions over time could also provide valuable insights into the working and development of children's minds, related to values, for social sciences. It could be a starting point for more, related or in-depth research into children's

relationship with values and/or with robots, as well as into children's learning abilities over time.

1.2. Research opportunity

In order to have a memory, robots need to be operating within a clear context and store the information specific to that context. In this case, that context is how children think or operate regarding different values. It is likely difficult or cumbersome to explicitly talk with children about values, since this involves talking about larger behavioral patterns and broad concepts. To overcome this, the conversations between robot and child are kept simple, recognizable and visual. However, this means that there is limited information to be gained from the verbal interaction between robot and child. A next step towards a more complete interpretation of the children's reaction to the value choices is to infer which topics provoke a reaction from them, spike their interest or seem memorable from non-verbal communication, such as body language, facial expressions and eye gaze.

Conversational robots are still a relatively new technology, so the information that is available about the effect that human-robot interactions (HRIs) have on humans is still in a developing phase. The effects of child-robot interactions (cHRI) on children is an even smaller data pool (Martelo & Villaronga, 2017). Any study into the effects and effectiveness of cHRI can therefore contribute to new insights in this relatively unexplored field.

Nevertheless, the multi-modality of human conversational behaviour during HRIs has previously been exploited in research into emotion recognition (Hong et al., 2021), engagement (Rich et al., 2010) and attention (Lemaignan et al., 2016). However, it has rarely been linked to what people remember. The ePartners4All project aims to develop an artificial conversational memory for a robot. It would be beneficial for the robot to not only store verbally provided data, but also reason over the child's state, i.e. memory. Namely, the robot could select relevant topics of conversation with the child, based on what the child remembers. Currently, there is some, although limited, literature available on estimating what adults remember from a conversation (Tsfasman et al., 2022). Even more scarce is the research on children's conversational memory (Stolzenberg et al., 2018). Research on children's conversational memory *during cHRI* is thusfar non-existent. Therefore, the knowledge of what children are likely to remember from a cHRI could provide a lot of new and meaningful insight that could be applied to the development of robots. More specifically, it could make sure that robots can engage with children in more meaningful interactions over longer periods of time.

Robots could exploit the knowledge of what a child would remember from a conversation by remembering or revisiting the same moment and thus creating a stronger bond between robot and child (Reese & Brown, 2000). Alternatively, robots could revisit parts that are not likely to be remembered by children to expand or enrich the children's memory of the interaction. In the value-choice setting at hand, it would be interesting to see how this multi-modal data relates to children's value preferences and decisions. Given the robot's ability to remember, it could keep track of a long-term model of moments that a child remembers. Such a model could predict or, at least, provide insight into the children's interests in the different values or moral dilemmas.

From a broader point of view, knowing what children remember from cHRI, and translating this to what they find important, could be used in robot teachers, to automatically tune the teaching content to be discussed for each child separately. In addition, a robot could use accumulated data of this type for an internal feedback loop. This feedback could suggest the adjustment of certain moments in cHRI design to make the interaction more memorable generally.

Famously, the eyes have been dubbed 'the window to the soul'. Currently, eye gaze can be tracked accurately with off-the-shelves tracking algorithms on long-distance ($> 20cm$) camera data (Hutt et al., 2019). This is a promising development that allows for insights into humans' inner processes without invasive or costly hardware. Within the conversational context, eye gaze tracking has been linked mostly to mental states like engagement (D'Mello et al., 2012; Hutt et al., 2017; Rich et al., 2010). This previous research has proven that eye gaze tracking is a low-cost, reliable method for drawing conclusions about someone's inner state (Hutt et al., 2019).

As such, it can be expected that this relatively easily acquired eye gaze data can also be linked to memory processes, such as remembering a specific moment. This has been researched before with adults (Tsfasman et al., 2022), but not with children, and especially not in the cHRI setting. This research will be the first of its kind to relate children's eye gaze patterns to whether or not they will remember moments during a cHRI.

1.3. Research scope

It has now been established that a research into the relationship between children's eye gaze during cHRI and the likelihood of the children remembering moments during this interaction is scientifically insightful and provides opportunities for societal applications. This section will define the scope of this study, which addresses this research topic. In particular, this study will aim to answer the following research question:

> How and to what extent can eye gaze tracking during cHRIs be exploited to identify moments that the child remembers from this interaction?

The so-called 'moments that the child remembers from a cHRI' will be referred to as **memorable moments** throughout this report and refers specifically to moments that children remember *immediately after* the cHRI in case. The research question is two-fold. It will be researched *how* to detect memorable moments by identifying appropriate machine learning models for the problem setting and identifying heuristics regarding eye gaze patterns related to memorable moments. It will be researched *to what extent* memorable moments during cHRI can be detected by critically reviewing the results and the research methods. Finally, the results will be linked back to their societal relevance and as such, suggestions and pointers will be made as to how this information can be used to improve cHRI, with respect to children's mental and physical health in classroom settings where robots can be deployed as teachers' aid. Included in this study is the data collection (and finding ground truth) of memorable moments and the investigation of the relationship between these moments and the children's eye gaze patterns. Not included in this study is the making of or exploration of an algorithm for children's eye gaze tracking.

1.4. Document outline

This thesis document will describe in detail all the steps that were taken to reach the answers to the question proposed in section 1.3 as adequately as possible. More precisely, in chapter 2 the background, current state of the art, and common practices related to this research will be further elaborated upon. This includes common practices in cHRI, the relation between gaze and inner state monitoring, prior research on detecting memorable moments and state of the art pattern recognition algorithms for this type of data. In chapter 3, the process of collecting the memorable moments data and gaze data, processing it and analysing it such that it leads to the envisioned results will be described. These final results will be presented in chapter 4, which are then discussed and placed into perspective in chapter 5. Possible improvements for this study and recommendations for the application will be explored in chapter 6. Finally, chapter 7 will summarize the findings and provide a resolution.

2

Background

In order for robots to be of aid in the longitudinal development of children, they should be able to find out, store, and retrieve information about these children on a level similar to the way that humans are able to assess each other's signals. This means that robots should possess the ability to derive useful (meta-)information from input. In particular, if robots had the capacity to assess what children remember, they can relate better to the child during interactions or reason over the child's interests. In this thesis, it is assessed how multi-modal data can be leveraged to improve the aptness of robots in child development, towards the ultimate goal of improving children's mental and physical health. More specifically, it will be researched if, and how, intelligent systems can detect memorable moments using children's external features. The study will utilize eye gaze data, acquired during conversations between robot and child, to predict whether the child will remember the content discussed in different moments. This chapter will describe previous work in this research area and identify the missing links that this study aims to address.

2.1. Conversations between robot and child

As technological advancements have progressed over the past years, virtual agents and robots have become more widespread and made their introductions in public spaces, for example the social robots JIBO (Hodson, 2014) and Furhat (Al Moubayed et al., 2012). Some examples are the NAO robot (Gouaillier et al., 2008), Among these public spaces are classrooms, which has lead to the introduction of interactions between children and robots, specifically robots like Pepper (Pandey & Gelin, 2018) and Thymio (Riedo et al., 2012). One humanoid robot, designed especially to work in education and with children, is the NAO robot[1](Gouaillier et al., 2008). The robot was first introduced commercially in 2008. Since then, it has been applied in educational settings in both short and long term situations. NAO allows for easy operation and adjustments and is therefore commonly used in classroom studies as well. The application of robots in classrooms has a lot of potential for added value. For example, it could address the growing problem of teacher shortage by giving children individual attention while the teacher can remain focused on the main activities (Edwards & Cheok, 2018). Robots in classrooms are also a way to provide personalized learning environments (Karna-Lin et al., 2006). Personalized learning environments have been shown to help in the development of learning skills, independent learning and general autonomy (Dabbagh & Kitsantas, 2012). In addition, it can lead to increased engagement by the students with the material (McLoughlin & Lee, 2010) Finally, by letting children engage with robots, children are more encouraged to engage with technology in general and pursue STEM areas (Saleiro et al., 2013).

Majgaard, 2015 monitored the use of NAO in classrooms for a year and this was shown to have positive effects on students' motivation and learning outcomes. Even though the robot, in the aforementioned study, proved to have this positive effect with merely directed dialog, other studies showed that this effect of the NAO on motivation and learning outcomes could be even stronger by leveraging visual cues (Baxter et al., 2017; Han et al., 2012). So, throughout time, the NAO has been enhanced with the incorporation of appropriate visual cues, such as eye contact and hand gestures (Csapo et al.,

[1]http://us.softbankrobotics.com/nao

2012; Kim et al., 2013). This personification, however, is generally based only on the robot-side of the interaction, without taking into account the conversational partner. More profound personalisation arises when robots can make adjustments to human behaviour and preferences through sensors and some type of memory (Dudzik et al., 2018; Saravanan et al., 2022).

2.2. Personalising robot behaviour

While there is still a lot to explore in the field of robot personalisation, large scientific and technological advancements have already been made in this area. It has been shown that the maintenance of an episodic memory in a conversational or social robot can allow the robot to evaluate verbal information in context (Paplu et al., 2022); increase social bonding and motivation (Campos et al., 2018; Saravanan et al., 2022); and save and exploit user-specific information such as demographic characteristics or interests (Paplu et al., 2022; Sekmen & Challa, 2013). As also indicated by Elvir et al., 2017 and Coronado et al., 2022, the next step towards more advanced personalisation would be to make use of the multi-modality of human behaviour in conversations. The majority of the aforementioned studies focuses mainly on personalisation with regard to verbal information, as provided by the user or some other context. For humans, analyzing each other's facial expressions/gaze happens instantaneously and subconsciously and it provides us with a lot of information (Jokinen, 2009). For example, body language can convey someone's emotions, predict whether someone is joking or sarcastic, and show the level of engagement in a conversation (Beattie, 2003). Especially the ability to assess user engagement has been a topic of interest in the field of human-robot interaction, since it may provide useful information in regards to the robot's performance both during and post interactions. Examples of different modalities of expression are the tone of voice, choice of words, eye gaze direction and facial expressions. The focus of this research, in particular, will be on eye gaze during a conversation.

2.3. Gaze detection

Before the availability, or even existence, of human-robot interactions, researchers were already interested in measuring and influencing the attention span of humans. As such, many studies have made attempts to detect and/or mitigate "mind wandering", which is defined as the unintentional attention shift from the current activity towards internal, unrelated thoughts, with the help of gaze data (Bixler & D'Mello, 2016). Mind wandering has been found to correlate with gaze through differences in fixations (eyes fixed in one location) and saccades (movements between fixations), where fixations were found to be more erratic during mind wandering (Reichle et al., 2010).

While this aforementioned study, similar to others from its time and before, was focused on attention during reading tasks, the studies by Hutt et al., 2017; Hutt et al., 2019; Hutt et al., 2021 are more closely related to the interactive setting. These exploit gaze patterns to detect mind wandering during lecture viewing and prove to generalize and perform better-than-chance and also intervene the mind wandering when it occurs. In these studies, the use of eye gaze tracking is proven to be a successful, cost-efficient and non-intrusive way of measuring humans' internal state. Using advanced technologies from the field of neuroscience would perhaps provide deeper insight into human though processes/brain activity, but are complicated to set up, costly and difficult to apply in real-life settings. Eye gaze tracking appears to be a commonly used and optimal trade-off between costs and accuracy in terms of measuring humans' internal state. Moreover, it has been shown that even an imperfect or incomplete set of gaze features usually renders satisfactory results (Hutt et al., 2017).

A growing number of studies is now also directed at human-robot interaction (Ben-Youssef et al., 2017; Coninx et al., 2016; Coronado et al., 2022; Dini et al., 2017; D'Mello et al., 2012; Lemaignan et al., 2016; Nakano & Ishii, 2010; Sekmen & Challa, 2013), where Ben-Youssef et al., 2017 shows that the vast majority of these works also exploit eye gaze as the primary indicator for attention. In addition, Nakano and Ishii, 2010 states that, when acquiring eye gaze data for this task, off-the-shelf eye trackers are sufficiently accurate. Therefore, following the state-of-the-art, eye gaze data will be used as the indicator for attention in this study. The working and performance of the eye tracker itself is not within the scope of this project and the software used is obtained from the IDIAP Research Institute[2] and is described by Siegfried and Odobez, 2022.

[2] https://www.idiap.ch/en

2.4. Moment recollection

Memory refers to the overall system and processes involved in acquiring, storing, and retrieving information. It involves the encoding of information into memory, its storage over time, and its retrieval when required (Baddeley, 2013). Experiments and tests on human memory usually require the test subjects to reproduce (recent) events up to their best abilities. Cued recall and free recall are two specific paradigms used to study memory retrieval (Roediger & Guynn, 1996). In cued recall, individuals are provided with specific cues or hints that support the retrieval of information from memory. For example, a person may be given a list of words and then provided with the initial letters of each word as cues to recall them. Free recall, on the other hand, involves the retrieval of information from memory without any specific cues or hints. Individuals are asked to retrieve as much information as they can from memory without any specific guidance. For example, in a free recall task, individuals may be asked to recall a list of words in any order. Whereas memories recollected through free recall might be the most accurate representation of actual memorable moments, it might be necessary to trigger memories with cued recall in this study.

Gathercole, 1998 showed that children, especially pre-teen children > 7 years old, store and retrieve memories in a similar way as adults, but in smaller quantities. This, in combination with children's shyness or insecurities in unknown (research) environments (Thomas & O'Kane, 1998), might lead to insufficient memorable moments identifications. To circumvent this outcome, presenting the children with cues from the cHRI, e.g. visual aid on screen, might boost their confidence or memory recollection abilities. Applying this strategy would mean that this study combines free recall and cued recall tasks in an unstructured manner, but the goal of the study is not to identify the performance of different recall tasks, but to identify memorable moments in general. Discriminating memorable moments as retrieved through free recall, cued recall or any other recall type will be left for future research.

2.5. Proof of concept

There are many reasons as to why one would want to approximate the inner state of a human during HRI. In general, different studies show that the more modalities of communication are analyzed, and the longer the same users are studied, the more accurate a robot can make its user models and predictions. Sekmen and Challa, 2013 shows this by featuring an extensive, long term learning process in which a robot in the health sector learns the preferences of patients and elderly through their body language. It combines users' historical (verbally communicated) decisions and facial data, much like this study, in addition to external factors for decision making such as temperature and time. The study showed that a system that would predict and react proactively to desires based on these inputs was preferred significantly over the non-learning version. While these are promising developments, and reason for more similar and in-depth studies, in most practical settings a robot has less input, e.g. only gaze data, and shorter interactions with people. For example, Dini et al., 2017 measures gaze during HRI to make a robot reason over what objects in space the human is aware of and where they might need help. The results showed that the framework, which combines eye tracking data with machine learning algorithms, was able to predict the users' actions and situation awareness with significant improvement over the baseline models. In these examples, the robot in question is (successfully) reasoning over the inner state of the human using eye gaze. However, the topic of interest in these studies is the user's 'current' awareness levels and the experiments were less concerned with the HRI's long term effects on the human, like how it affects their memory. On the contrary, D'Mello et al., 2012 developed a virtual tutoring system that uses students' gaze to measure their engagement, to enhance long-term learning. When the system detects high levels of disengagement, the tutoring system intervenes and tries to recapture the students' attention. As such, the researchers aim to provide a more efficient and effective way of teaching students topics that they retain on the long term. The study proved that students that used the intelligent tutor performed significantly better after the fact on the content that was studied, than the students that were given the non-adaptive version. This suggests that leveraging eye gaze data to reason over humans' inner states is not only statistically successful, but can also be applied practically with effects on a term longer than just in the moment. Nakano and Ishii, 2010 conducted a similar experiment, where a virtual agent was explaining the functionalities of different objects on screen and the agent would ask probing questions if users seemed to disengage. In their study they also dissected specific gaze patterns related to high engagement and low engagement, which will be explored more in section 2.6. Lemaignan et al., 2016 measured "with-me-ness" (Sharma et al., 2017),

a combination of engagement and attention metrics, using eye gaze during teaching tasks between a NAO robot and child. Even though the number of participants (6) in the study is too small to draw significant conclusions, this study did show promising results in the form of accurate translation of the gaze direction to the 'with-me-ness'. The experimental setup of Lemaignan et al., 2016 is quite similar to the one in this study, so we would expect similar good results, but with a more rounded and substantial dataset.

Generally, in the HRI domain, gaze data is most often correlated with engagement and attention. In this domain, engagement relates to the extent at which the human is involved and invested in the interaction with the robot. Attention refers to the focus of the human on a specific aspect of the interaction, e.g. a human can direct its attention to the robot itself or to the object that the conversation is about. Within the eye gaze context, the object of attention is also called 'visual focus of attention' (VFOA). Arguably, in order to remember the content or topic of a conversation, a human needs both: a minimum level of engagement as well as an appropriate or meaningful VFOA. So, while there is limited literature available on the relation between gaze patterns and memorable moments detection specifically, literature about engagement and attention detection may still provide useful heuristics.

2.6. Heuristics for memorable moments detection in cHRI

In their study that involved onscreen tasks guided with an onscreen agent, Nakano and Ishii, 2010 found that participants who were highly engaged in the conversation tended to have longer fixations on the agent's face and eyes, and more frequent saccades between the agent's face and eyes. The study also found that participants who were less engaged tended to have longer fixations on other parts of the screen, such as the background or other objects. So, gaze patterns that involve more attention to the agent's face and eyes are more indicative of high engagement in a human-agent conversation. The study proposes a model for estimating engagement levels based on these metrics and has created a system that probes questions when users seem disengaged. The model was proven to cause statistically higher engagement levels throughout the interactions when compared to a non-probing system. Since engagement detection is somewhat related to memorable moments detection, similar gaze patterns would be expected to emerge from this study. An important difference, however, is the use of a virtual agent versus the use of a robot. While the robot is likely initially more interesting to engage with, it lacks facial expressions, so this can be a cause for diverging results.

As mentioned before, Lemaignan et al., 2016 measures 'with-me-ness' during cHRI with a NAO robot. While the experiment setup of this research will be described in detail in chapter 3, it is worth noting that it is very similar to the one in Lemaignan et al., 2016. In both cases, the NAO robot stands across from the child on a table, a tablet is placed between the robot and the child, and the gaze recognition camera is positioned near the legs of the robot behind the tablet, so the data formats, and thus research possibilities, are quite similar. In order to obtain 'with-me-ness' as a value, Lemaignan et al., 2016 created predetermined tasks with predetermined visual targets (i.e. robot, tablet, etc) that would be classified as 'with me' in case the child looked at it during the task execution/timeframe. The model built from this data, when compared to the manually annotated ground truth, has very high accuracy. It should be noted, however, that the very definition of 'with-me-ness' is subjective to the researchers and not verified with the users during or after the experiment. This is a stark difference between the studies, since in this study it is the participants that identify the moments that they remember. However, the study shows, through correctly predicting the visual focus, that children who are engaged (or rather, 'with-me') gaze at the visual object in space that follows logically from the topic of conversation. This is especially relevant information because of the common use of an assisting tablet during the interaction. Namely, this information would suggest that the participants engage with the tablet (or screen) during the moments for which it serves a function. This is in line with the findings in Jermann and Nüssli, 2012, in which two collaborators are shown to work better together when they have a shared visual focus of attention that is in line with the topic of conversation.

Finally, as an additional source of heuristics, especially since this study uses a robot (contrary to the common use of virtual agents in other studies), it is worth looking at relevant research in the domain of human-human interaction. Tsfasman et al., 2022 conducted a study in which a system is developed that predicts memorable moments in un-directed multi-party (4 person) conversations. The key findings in this study with respect to gaze are that, at moments of high memorability, the participants were less likely to look at each other and more likely to look at the same visual target. It should be noted that

the NAO robot, used in the experiment of this study, will not be equipped to change gaze and will remain visually focused on the conversational partner (a type of behaviour that has been shown to decrease engagement levels of the user (N. Wang & Gratch, 2010)). It is therefore worth looking extra into behaviour related to mutual facial gaze between conversational partners. According to Rich et al., 2010, mutual facial gaze between people is an indicator for both people that they intend to maintain engaged in the conversation. In addition, mutual facial gaze is common at the end of a speaking turn. It has also been shown that people show less mutual gaze when disclosing intimate information (Kang et al., 2012). On the contrary, high levels of mutual gaze has also been linked to high levels of trust (Normoyle et al., 2013). Any of these observations could be linked to the memorability of a moment and will be taken up for discussion.

It is important to keep in mind, however, that children are likely to have different reactions to a robot than adults. This may manifest in them having a more curious or perhaps scared attitude towards the robot, and they are more likely to attribute human qualities to it, especially younger (4-8 year old) children (Burdett et al., 2022). There has been very little research done in the differences in gaze behaviour between adults and children, so it's hard to translate the heuristics for adults to children. There are some studies that did research the difference between adults and children in a non-interactive task-execution setting and in these cases there were no statistical differences found in gaze behaviour (Blythe et al., 2009; Mackworth & Bruner, 2009), so in this study it is generally assumed the adult gaze heuristics to be applicable to children.

2.7. Hypothesis

Based on the heuristics regarding human gaze behaviour during conversations, the following hypotheses are drawn up. The goal of the hypotheses is to identify common patterns or themes in gaze behaviour for memorability and to possibly link this to existing knowledge on gaze behaviour in relation to memorability or other inner processes.

1. Children who are likely to remember a moment during the cHRI exhibit different gaze-time distributions over the different visual targets, compared to children who are not likely to remember a moment.

2. Gaze patterns at the start of a scenario, i.e. when the scenario topic/content is introduced (≤ 35s) and new information is incoming, are more indicative of the scenario being memorable or not than those later in the scenario discussion.

3. Children are less likely to remember scenarios if they are indecisive regarding the choice to be made during these scenarios, compared to children who appear decisive.

4. Children who self-identified their reason for remembering a scenario to be its (cognitive) ease spend more time looking at 'other' and 'robot' and less time looking at 'screen'.

5. Children who can reproduce 'high quality' recollections of a scenario spend more time looking at ('studying') the screen during these scenarios.

6. Children who indicate that they remember a scenario because of how it made them feel are looking more at 'robot' and 'other' compared to children who indicate to remember a scenario because of the content, who, in turn, are looking more at 'screen'.

2.8. Models for memorable moments detection

The problem of 'detecting memorable moments' can be classified as a pattern recognition problem. Because of the general lack of available data and heuristics that relate gaze patterns to memorable moments, machine learning models provide an excellent method to infer these patterns computationally. By its very definition, gaze patterns are defined by the succession of gazes in different directions, i.e. a time series. This section will provide a brief overview of the technological developments and current state of the art in the domain of time series classification.

2.8.1. Time series classification: an overview

There is a large variety of algorithms available for the classification of time series data. These algorithms can be roughly divided into the following categories: distance-based, interval-based, dictionary-based, frequency-based, shapelet-based, convolution-based, deep learning-based, and ensemble methods (Bagnall et al., 2016). In order to justify the right model of choice for this study, this section will explain the basic architectures behind each category.

Distance-based An example of a distance-based machine learning algorithm is k-nearest neighbors. This algorithm can be adapted to fit the time series problem by replacing the Euclidean distance metric with the dynamic time warping metric. Dynamic time warping is a distance measure for finding the similarity between two time series, while accounting for the fact that they may not align exactly in time, speed, or length (Müller, 2007). This combination, KNN-DTW, is commonly used as a benchmark for evaluating time series classification algorithms because it is simple, robust, and does not require extensive hyperparameter tuning. While useful, KNN-DTW requires a lot of space and time to compute, because it compares each object with all the other objects in the training set during classification. Further, KNN-DTW provides limited information about why a series was assigned to a certain class (Ruiz et al., 2021).

Interval-based Interval-based algorithms split the time series into random intervals, with random start positions and random lengths. Then, summary statistics over each interval are computed and put into feature vectors, that are used to train a classical machine learning model. The full series are classified according to a majority vote of all trained models (Rodríguez et al., 2005). An example of this structure is the canonical interval forest (CIF), which is comprised of a number of decision trees. CIF has been shown to be computationally efficient, interpretative and to outperform KNN-DTW in terms of accuracy (Middlehurst et al., 2020). Middlehurst, Large, Flynn, et al., 2021 proposed some minor improvements over the CIF model and this updated, current standard model goes by DrCIF.

Frequency-based Frequency-based classifiers are similar to the interval-based ones, except that they extract spectral features from the series, instead of summary statistics. One simple classifier is trained per interval and the extracted features from each interval are concatenated to form a new dataset. An example of such an algorithms is the Random Interval Spectral Ensemble (RISE) (Flynn et al., 2019). In order to limit the computational costs of RISE, a simple time or other resource constraint can be set and the algorithm will simply keep building trees until the limit. However, in the case of long time series, this may result in a small ensemble (few trees), because it is computationally expensive to go over one time series (Flynn et al., 2019).

Dictionary-based Dictionary-based classifiers use sliding windows of a pre-determined length and convert the time series data in each window into a so-called 'word'. A dictionary of these words is constructed as the window slides, while keeping a count of each word's frequency. On the resulting histograms extracted from the time series, any classifier can be trained. The current state-of-the-art dictionary-based algorithm is Temporal Dictionary Ensemble (Middlehurst, Large, Cawley, et al., 2021). The TDE algorithm is fast, robust, interpretative and among one of the top time series classifiers according to the experiments in Middlehurst, Large, Cawley, et al., 2021.

Shapelet-based Shapelet-based classifiers search for consecutive sub-sequences of the time series that have discriminatory power. These shapelets determine the likeliness of different classes. Given n extracted shapelets, each feature vector of the new dataset will have n dimensions, one for each distance to each shapelet. Any classification algorithm can be applied to the new dataset. Depending on the implementation, the shapelet selection process can be computationally expensive, but the currently best ranked shapelet algorithm, the shapelet transform classifier (STC), retrieves all shapelets in a single pass. STC combines them using a rotation forest classifier (Bagnall, Flynn, Large, Lines, et al., 2020; Lines et al., 2012), which was proven to be the best classifier for continuous data (such as distances) in Bagnall, Flynn, Large, Line, et al., 2020.

Convolution-based Another specific algorithm that has been proven to perform very well (Dhariyal et al., 2020; Ruiz et al., 2021) is ROCKET. ROCKET is a simple linear classifier based on random convolutional kernels (Dempster et al., 2020). ROCKET is easy to use, as there is no need for endless hyperparameter tuning and provides high classification accuracy at minimal cost. On the downside, it lacks interpretability.

Ensemble-based Finally, the HIVE-COTE algorithm is a meta-ensemble based on several of the classifier-types mentioned before. To be more precise, HIVE-COTE predictions are a weighted average of predictions produced by its members: STC, BOSS, Time Series Forest (an interval-based classifier), and RISE (Lines et al., 2016). Despite being more reliable, the first stable version of HIVE-COTE (version 1.0) was not significantly better in terms of accuracy than other high-performing time series classifiers (Bagnall, Flynn, Large, Lines, et al., 2020). However, recently a newer version was introduced, HIVE-COTE 2.0, which has been proven to be significantly more accurate on average than the current state of the art on 112 univariate UCR archive datasets (Dau et al., 2019) and 26 multivariate UEA archive datasets (Bagnall et al., 2018; Middlehurst, Large, Flynn, et al., 2021). As a downside, HIVE-COTE is less interpretable than its members individually and is computationally expensive.

Deep learning-based Deep learning models have made their introduction to problems of almost all types, due to their ability to handle complex problems at high success rates As such, deep learning has also made its way into the field of time series classification. Whereas a method like HIVE-COTE struggles with scalability issues, deep learning methods can perform on par in terms of accuracy, while handling much larger datasets and training much faster (Ruiz et al., 2021). An example of a deep learning algorithm for time series classification is InceptionTime. This classifier is an ensemble of deep Convolutional Neural Network (CNN) models based on the Inception deep learning model for image recognition and performs similar to HIVE-COTE at lower computational cost, given the availability and use of a GPU (Fawaz et al., 2020; Szegedy et al., 2016). As with all deep learning methods, the computation functions as a black box and is very hard to retrace and interpret. For this reason, deep learning approaches algorithms are not further included in this study.

2.8.2. Multivariate time series classification

For the data collection in this study, the acquired sensor data is multi-dimensional (multivariate), since it contains gaze attention flags for different visual targets in the room (for more details see chapter 3). Even though multivariate data is generally more common than univariate, much less attention has been given to the multivariate time series classification problem, compared to univariate time series classification (Ruiz et al., 2021). As a result, algorithms for the multivariate case are less available or merely a built-in translation from multivariate to univariate. This can be achieved through randomly sampling dimensions or using other dimensionality reduction methods, but this leads to data loss. Another option is to concatenate the dimensions, but this leads to the potential loss of information regarding dimension interactivity per timestep. According to a recent multivariate time series classification survey, scalability is a big issue in most state-of-the-art multivariate time series classifiers (Dhariyal et al., 2020). The same survey also appointed ROCKET as the best performing algorithm, despite its failure to complete some problems (Dhariyal et al., 2020). A similar, more extensive survey also compared different multivariate time series algorithms and found that the distance-based benchmark was still hard to beat for many of these algorithms (Ruiz et al., 2021). Nevertheless, ROCKET was found to be the best overall, given its speed and high accuracy. CIF was also placed in the top 3 and perhaps this is due to the fact that multivariate ROCKET and CIF both make use of some dimension dependent feature extraction (Ruiz et al., 2021). The improvement of bespoke multivariate CIF over simple dimensionality reduction methods suggests that there is a lot to gain from the adaptation to the multivariate case and if this were to be introduced in more sub-components of ensemble methods like HIVE-COTE, these ensembles would likely also improve instead of fall behind in this setting (Ruiz et al., 2021). As a final remark, the field of multivariate time series classification has been making fast, recent developments. This means that some of the models mentioned before in subsection 2.8.1 also have bespoke multivariate implementations now, but they have not been thoroughly reviewed in literature or surveys, which makes it all the more interesting to compare their performance in this study.

2.8.3. Models of choice

Bagnall, Flynn, Large, Lines, et al., 2020 argue that HIVE-COTE is the best default option, i.e. best in the case of little domain knowledge, since it is an ensemble of classifiers built on different representations and achieves high accuracy. However, HIVE-COTE has no bespoke multivariate implementation, due to most of its members not having this implemented, so its accuracy drops in the multivariate setting. In addition, HIVE-COTE is computationally expensive and this is arguably unnecessary, given the fact that there is domain knowledge available, described in section 2.6. ROCKET, on the other hand, performs well in the multivariate case and is very efficient. ROCKET is also widely applicable for any problem setting off-the-shelf and requires little to no hyperparameter tuning. Therefore, ROCKET will serve as benchmark model for this study. As a downside, ROCKET lacks interpretability. For this reason, it was decided to also include classifiers based on more local features as well. Whereas shapelet-based classifiers will be better when the best data feature is the presence or absence of a pattern, dictionary/frequency/interval-based methods will be better when you can discriminate using the frequency of a pattern (Large et al., 2018). Given the heuristics from section 2.6, one might expect to discriminate between gaze patterns of memorable moments and those of non-memorable moments by the occurence or number of occurrences of sub-patterns within the time series, like a sequence/fixation of looking at the robot's face and away (saccade). Therefore, each of these type of classifiers could provide good or insightful results. It was decided to use STC for the shapelet-based method and TDE for the dictionary-based method. There is no known state-of-the-art implementation of a frequency-based method for the multivariate case, so this will be left for future research. The CIF algorithm has been proven to perform well in the multivariate case and is therefore chosen to be the final participating model in this study, to represent the interval-based method.

3

Method

To answer the research question, *"how and to what extent can eye gaze tracking during cHRIs be exploited to identify moments that the child remembers from this interaction?"*, there should be eye gaze data of cHRIs and ground truth data regarding memorable moments during this cHRI. Then, intelligent algorithms and statistical analyses can be used to draw conclusions based on this data. In order to obtain this ground truth data, there needs to be an experiment that involves a cHRI and a recall exercise, followed by appropriate data processing and formatting. The details behind the design choices that were made to successfully complete this study are laid out in section 3.1 and the hardware and software setup to bring the design to life are described in section 3.2. Relevant demographic information regarding the participating children can be found in section 3.3 and the tasks they have to complete as part of this experiment are described in section 3.4. The steps that were taken to obtain workable, machine-readable data from the experiments are described in section 3.5. Finally, the statistical methods used to analyze the data are described in section 3.6.

3.1. Design

Given the lack of existing literature regarding the relation between eye gaze and children's conversational memory (of cHRI), the research in this study is of exploratory nature. In particular, the study follows a correlational design that tests the naturally occurring relationship between two variables: children's eye gaze during cHRI and the memorability of value-based decision-making moments.

3.1.1. Variable 1: Eye gaze

The first variable is children's eye gaze during a cHRI in which the children are asked to make value-choices. As it is stated in the research question, information regrading the children's eye gaze is integral to this study. However, taking eye gaze direction unprocessed, i.e. as a point or vector into space, introduces a lot of noise. This is due to the fact that the gaze tracking setup is not equipped for that level of precision. Using the raw data will assume the gaze calculation to be more precise than it actually is and this could lead to overestimation of the results. In addition, the cHRIs are relatively short and task-oriented, which should make it easy to set up or identify relevant visual targets.

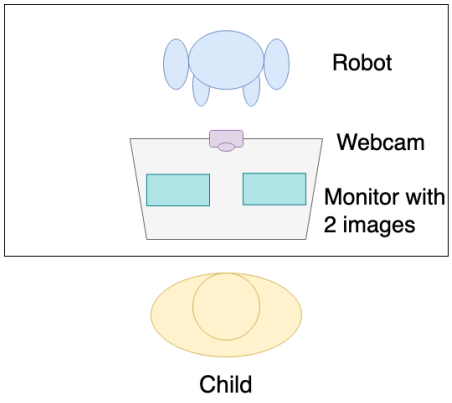


Figure 3.1: Top view of the experiment setup of the robot, child, screen and gaze camera.

By design of the experiment, the children have to choose between two options, guided by values, in fictional everyday scenarios. It was decided to display a visual on the screen, during each explanation and discussion of a scenario, with two images that each represent one of the available choices that the child can make. The full setup can be seen schematically in Figure 3.1 and photographically in Figure 3.5. An example visual is displayed in Figure 3.2. The images serve two

functions. Firstly, they help the child to understand the robot's explanation of the scenario and the options. During this explanation, the robot talks for a relatively long time and the child might get distracted or lose track. The addition of the visual aid should help them conceptualize the content. Secondly, the literature presented in chapter 2 (Lemaignan et al., 2016 and Jermann and Nüssli, 2012) demonstrated the relevance of visual targets during (human-robot) conversations, and that there is information to be gained from the eye gaze when having a common visual focus of attention between conversational partners. Adding the images on screen provides extra visual points of interest that can be leveraged to determine memorable moments. In addition, the colorful and interesting visuals may increase the memorability of that scenario or that moment and thus might increase the success rate of the experiment.

In total, the relevant gaze targets identified in this experiment are the conversational partner, i.e. the robot, the left side of the screen, displaying the first choice, the right side of the screen, displaying the second choice, and finally any point in space that does not fall within these categories. Children's gaze behaviour regarding these visual targets should already provide a lot of insight into children's general inner processes like deep thinking or conversational turn-taking. Related literature has had successful outcomes regarding behavior prediction using eye gaze with similar designs Lemaignan et al., 2016. Nevertheless, taking these 4 visual targets instead of raw eye gaze direction also leads to information loss. For example, if a child is looking at the left side of the screen and has a lot of frequent small gaze changes (a lot of saccades and short fixations) within that area, the current approach would register that as one long fixation on the left side of the screen, without saccades. This information loss is accepted in this study, as the main visual target in the example remains the left side of the screen, but the nuance is taken into account during analysis and results discussion.

Figure 3.2: One of the images that was displayed onscreen during the cHRI to aid the child's decision making progress. Throughout all images, the main subject ("you") is displayed as the girl with the blue shirt. The scenario for these visuals is as follows: "You are playing tag at school during recess. You are 'it' and you tagged a classmate, but they deny that you touched them. What do you do?" The left image represents choice: "You start a discussion to insist that you did tag them." The right image represents choice: "You accept their denial and continue being the tagger."

3.1.2. Variable 2: Memorability of moments during a cHRI

The second variable is the memorability of moments during the cHRI. Since the children are likely not able to remember or explain details beyond the topic of conversation and the cHRI is conveniently structured in different topics, it is chosen to consider each scenario discussion as a single 'moment'. Such a moment can either be memorable, i.e. the child remembers and reproduces it, or not, i.e. the child does not mention the scenario. This memorability is measured through an interview with the children immediately after the cHRI.

Order of tasks There are several reasons as to why it is chosen to conduct the post-interaction interviews straight after the cHRI. Firstly, the participants are all spending their time together at the robot camp doing other activities in between the cHRIs. This means that they have plenty of opportunity to discuss their experiences and thus influence each other and their personal memories of what happened during the cHRI. Secondly, the participants are spending their whole day interacting with robots and other children that they don't know, as part of a robot camp, and are getting a lot of new information and stimuli. This can be a lot to process, especially for the younger children, and can interfere with memories of the cHRI related to this experiment, such that the children remember less or confuse memories of the day. These problems are mitigated by doing the post-interaction interviews directly after the cHRI. As a consequence, however, it is hard to conclude from this study what moments children actually remember from cHRI on a longer term, like several hours, days or weeks.

Interview style As described in chapter 2, there are different paradigms for triggering memory retrieval. Moments reproduced by free recall could be seen as the most memorable, as they are spontaneously reproduced and not triggered by external factors. Therefore, this type of recall is sought after in this study. These moments could be obtained in an interview through a simple, open-ended question like 'what do you remember?' However, as was also laid out in chapter 2 (Roediger & Guynn, 1996), this style of recall might be too vague and non-committal for children to respond to. Therefore, when the child indicates that they are not remembering anything (else), they are prompted with cues to aid the memory retrieval process (i.e. cued recall). More specifically, these cues are the images that were presented on screen during the cHRI (e.g. Figure 3.2). This should give the children more confidence in and grip on the recall task at hand, hereby increasing the overall recall rate. Even though mixing different types of recall is not standard practice, at least for research with adult participants, for this study it has added value as it provides more security for the participants and more opportunities for the researchers. Moreover, at its core, cued recall and free recall address the same task, which is the recollection of past moments.

For the interviewers, the most important instruction is to always keep the conversation in line with the goal of the experiment: to find out what the children remember from their interaction with the robot and why. Figuring out *what* the children remember leads to the identification of a ground truth 'memorable moment'. Asking the participants to retrieve their memories is a clear objective and it should be attainable with simple, direct questions. A script has been prepared that explicitly states the questions that should be asked and in what order. However, there is a large variety in the children's openness, willingness to participate and ability to remember things and express themselves. For this reason, the data collection process will benefit from a semi-structured interview approach. This means that, while the interviewer follows the script by default, the interviewer can also deviate from the script if the situation calls for it. The script is therefore used more as a guideline rather than a 'straitjacket'. It is also for this reason, that the interview is held verbally, as opposed to as a written questionnaire, as it allows for more flexibility and opportunities to explain any ambiguities. Another reason for the verbal interview is the fact that some children are very young and might struggle with writing (large pieces of text in limited time). This would perhaps lead to a lack of details and difficulties with interpreting. The full post-interaction interview instructions/briefing for the interviewers can be found in Appendix A.

Thematic analysis For a deeper analysis and understanding of the children's conversational memory and their correlation to eye gaze patterns, it will be researched *why* the children remember certain moments and not others. The question "why do you remember that?" will be asked to the participants, after they manage to reproduce a moment. However, providing the answer to this question may require a level of reasoning and understanding over one's own mind that is difficult for the young children to attain (resulting in countless "I don't know"s as answers). Nevertheless, the reason for remembering can also be deduced from the children's subjective memory reproduction, since the objective truth in regards to the scenario content is known. Through a proven scientific method, called reflexive thematic analysis (Braun & Clarke, 2006, 2019), this subjectivity is captured and grouped into themes (or: labels). Reflexive thematic analysis involves a recursive and reflective process of identifying and interpreting patterns of meaning within the data, which can inform the matching implications. This approach emphasizes the importance of the researcher's subjectivity and reflexivity in the analysis process. To use reflexive thematic analysis to extract themes from textual interview data, a common use-case, the researcher follows a multi-step process that involves:

1. familiarizing yourself with the data through repeated readings and noting initial impressions and thoughts;
2. generating initial codes or labels that capture key concepts and ideas in the data;
3. grouping these codes into potential themes;
4. reviewing and refining these themes to ensure they are coherent, meaningful, and relevant to the research question;
5. mapping the relationships between the themes and the broader context of the research; and
6. putting it all together and writing it in a report.

- Braun and Clarke, 2012

label	requirement
low quality	The participant indicates to remember a scenario and is able to identify which scenario by any means or reference.
medium quality	All criteria above.
	The participant correctly describes the scenario's problem statement.
	At least one of the following: •The participant correctly describes the two options provided as response to the problem. •The participant recalls which decision they made.
high quality	All criteria above.
	At least two of the following: •The participant describes items or characters as they were described or displayed in the scenario (e.g. 'the girl was with her friends'). •The participant describes the emotional state or appearance of characters in the scenario (e.g. 'the teacher was angry'). •The participant mentions a specific word, phrase or item that stood out (e.g. 'sandwich', 'flute' or 'two left arms'). •The participant expresses their opinion about the scenario or decision (e.g. 'it was sad').

Table 3.1: The rules for classifying remembered moments as low, medium or high quality.

The labelling process will not only be based on thematic analysis, but also on heuristics for labelling humans' self-identified reasons for remembering conversations in literature. Tsfasman et al., 2022 have established 7 labels and 13 sublabels of exactly this kind. These labels are good suggestions of what type of information to look for and how this could be grouped. These labels are reproduced in Table B.1 in Appendix B for convenience. It should be noted, however, that there are some differences in the experimental settings between this study and the research by Tsfasman et al., 2022 that could give rise to divergent labelling. To be more precise, in Tsfasman et al., 2022 humans are interacting only with other humans (i.e. no robot), the setting is multi-party (i.e. $3-6$ conversational participants instead of 2), the participants are all adults (i.e. of age >18 instead of $7-11$), the conversations are longer (i.e. ± 45 minutes instead of $20-30$) and the conversations are not guided or scripted beyond the determination of a broad topic, e.g. the COVID-19 pandemic. The labels in Table B.1 are therefore used merely as a starting point or basis for the thematic analysis that will follow, rather than as a fixed set of rules to copy.

Rule-based labelling Possibly, one child manages to perfectly reproduce the problem statement of a scenario as it was told by the robot and another child only manages to identify a scenario with a single word or by only pointing at a picture. To distinguish between these levels of reproduction, the memorable moments are rated on the *memory quality*. Making this distinction could provide insight into whether eye gaze patterns are also correlated with the memory quality or level of detail that a child can produce regarding this moment. In addition, this labelling in combination with the other labelling, for the reason for remembering, might give some insights into what reasons or methods for remembering are more effective. To mitigate researcher subjectivity influencing the memory quality rating, the rating follows a strict set of rules. The rules were drawn up by the researcher during the thematic analyses, based on observations made during repeated analysis of the post-interaction interviews. These rules are laid out in Table 3.1.

Quality of labelling In order to ensure the quality of the data labelling, the data was completely anonymized before the process. The full dataset was labelled by the main research owner. To verify the labels' correctness and interpretability, a second researcher independently labelled a subset of the data and the resulting labels were compared to those of the main researcher. The second researcher reviewed 25% of the full dataset and the labels matched for 89% with those of the main researcher. This percentage was deemed sufficiently high to maintain the labels as provided by the main researcher for the remainder of the study.

3.2. Setup

3.2.1. cHRI

Hardware The hardware setup has been presented schematically in Figure 3.3. During the cHRI, the NAO robot ($60cm$ tall) stands or sits on a table, facing the child that sits on a chair across from it. In between the child and the robot is a big screen (19 inch diagonal/$16:9$ ratio/$1920x1080$ pixels). Just on top of the screen, which is tilted roughly 15 degrees towards the child to facilitate a clearly visible and distinguishable line of sight and elevated position, is a camera that has the purpose of recording and tracking the child's eye gaze. The gaze tracking camera records the child in Full HD (1920x1080p). A similar, additional camera is placed diagonally behind the child as a control measure and records the back of the child and the screen. A separate microphone records the audio of the cHRI with 16 bits per sample and a sample rate of 44.1kHz. The post-interaction interviews were audio-recorded using microphones with the same specifications.

Visual aid Since the child is requested to decide between two options, two images will appear on screen as one visual during the discussion of the corresponding scenario. The visual focus of attention of the children will be captured for each image separately.The algorithm manufacturer indicated that the gaze tracking algorithm works best for distinguishing a maximum of $4-6$ equal parts of the screen. With this information in mind, and to maximize separability between different gaze targets, it was decided to conceptually divide the screen into 6 equal parts: 3 parts horizontally and 2 parts vertically, so as to create 6 distinguishable parts as shown by the black lines in Figure 3.4. The two visuals that should be onscreen per each discussed scenario during the cHRI are then placed in the leftmost and rightmost horizontal divisions, leaving the third, center area empty for maximum separability. The placement of the images on screen are indicated by the red lines in Figure 3.4. In the resulting dataset, the children will be said to look at the left image if the gaze coordinate falls in the left area, indicated by the blue lines in Figure 3.4, and the same for the right image and the right blue area. An image with two example visuals is shown in Figure 3.2.

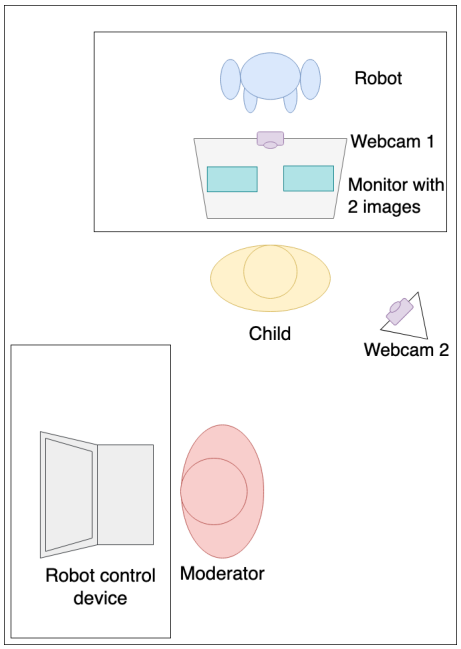


Figure 3.3: Top view of the experiment setup, displaying all essential components.

3.2.2. Software

Data processing In order to convert the raw audio and video data from the experiments to workable, machine-readable data, several programs have been used. The gaze extraction algorithm is a Linux executable, that was run on an Ubuntu 18.04 LTS system. For any playback or audio manipulation functions, the local program Audacity was used. Automatic speech recognition was locally executed by Whisper. OpenAI's Whisper is the best choice for ASR, since it is known to perform competitively among big cloud-based solutions provided by e.g. Google and Amazon and it allows users to download and run it locally, thus not compromising the data security. Merging, linking, processing and formatting data was done in Python, using the DataFrames structure provided by the package `pandas`.

Classification The machine learning models are trained in Python, because of its vast array of readily available libraries for data analysis and machine learning. A commonly used machine learning library is `scikit-learn`, which offers a range of widespread machine learning algorithms. By default, `scikit-learn` has limited support for time series data. However, `aeon` is a recently published (initial release March 2023, release used in this study May 2023) framework for time series tasks that extends the `scikit-learn` interface. `aeon` provides a large library of time series algorithms and interfaces with other time series packages to provide a unified framework for algorithm comparison. ROCKET, DrCIF, STC, TDE and HIVE-COTE v2 are all present in `aeon`'s library and are interfaced from `aeon`'s

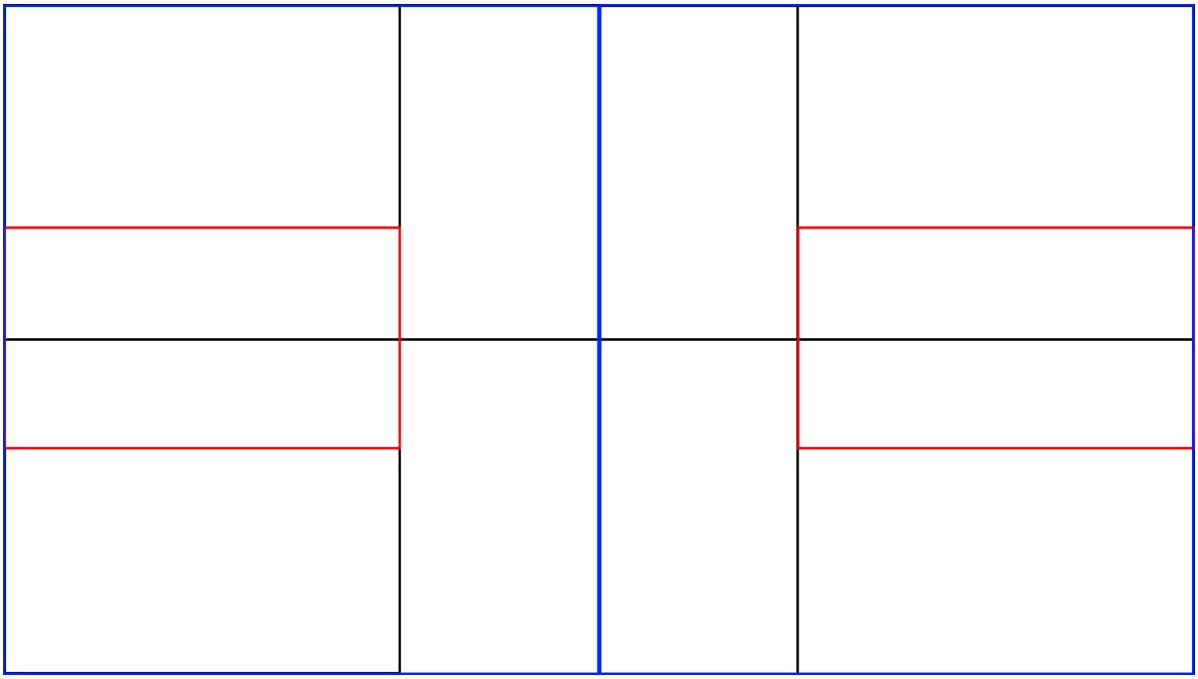


Figure 3.4: The conceptual division of the screen to maximize the separability of the two gaze targets on screen. The black lines indicate the maximum division precision of the gaze tracking algorithm in 6 separate areas. The red lines indicate the placement of the two images on screen. The blue lines indicate the screen areas that will be classified as looking on the left side of the screen and looking on the right side of the screen.

python API in this study.

Statistical analyses For the manual, statistical analyses, a range of different products and programs was used. Part of the analyses were done in Python in Visual Studio Code on macOS, using the package `pandas` for data handling, `matplotlib.pyplot` and `seaborn` for plotting and `stats.scipy` for statistical tests. Python is not fully optimised for statistics and `stats.scipy` does not offer the full range of existing statistical tests, contrary to R, which is designed for statistical computing. Therefore, part of the analyses were done in R in RStudio on macOS, using the package `vegan`.

3.2.3. Data protocol

In order to fulfill the study goals, the cHRI is being recorded visually from 2 angles, the back/side and the front, in which the participants are clearly identifiable. In addition, the audio of the cHRI, as well as the audio of the post-interaction interview, is being recorded. The video and audio data is stored without personally identifiable information; only by participant alias and time of the data collection. In order to ensure the participants' privacy and safety, this data is stored safely and securely. During the data collection process, the data is being saved to encrypted (external) hard disks. After the data collection, the data is transferred to encrypted servers with secure access protocols. The processing of the post-interaction interview data, which is specific to this study only, is executed locally on encrypted hard disks and the local copies are removed immediately after the data processing finishes. The TU Delft Board of Ethics approved this project and the corresponding Data Management Plan and Data Protection Impact Assessment, which also adhere to the European GDPR.

3.3. Participants

A total of 30 children participated in this experiment. 18 of the participants are male and 12 are female. Their age ranges from $7-11$ years old, with a mean of 8.5 years ($SD = 1.2y$). Each participant takes part in at least one and at most 2 separate cHRIs, depending on how many scenarios are discussed in one session (either 4 or 8). In case the participant takes part in 2 separate sessions of 4 scenarios, these sessions are recorded on separate days.

The cHRI is held in dutch, so the participants should be fluent in dutch. The experiment is held during school holidays and is part of a robot camp, organised by RobotWise[1] in Amstelveen, the Netherlands. The participants, or rather their legal guardian(s), signed up for this camp voluntarily and participation in the research is optional and also on a voluntary basis. The results of this study might be influenced by the fact that these children voluntarily signed up for a robot camp and are therefore likely to have a positive attitude towards robots in general. In addition, the participation in the camp came at a price and the location was in a municipality with one of the highest average incomes in The Netherlands ("Ranglijst van het hoogste en laagste gemiddelde inkomen per inwoner van de gemeenten in Nederland (bijgewerkt 2023!)", 2023). This might result in a bias in the participants' socio-economic backgrounds, with respect to the rest of society, which can influence their education levels, affinity with technology and cultural manners.

3.4. Task

The following section will describe the tasks that the children had to do in order to complete their participation in the experiment. These can be separated into two main parts, the cHRI and the post-interaction interview.

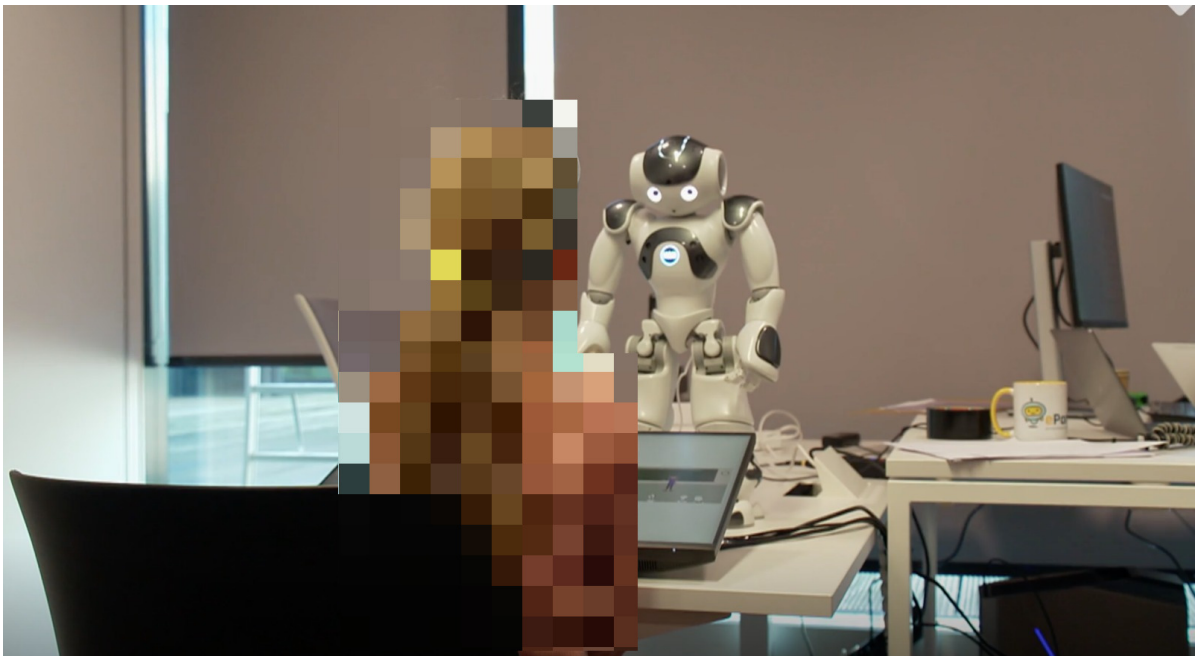


Figure 3.5: The cHRI setup with a participant (anonymized) facing the NAO robot and the tilted screen.

3.4.1. Child-Robot Interaction

During the interaction, the robot follows a scripted dialogue in which the goal is to gather information regarding the child's inner values. As defined by Schwartz, 2006, these values are achievement, benevolence, self-direction and conformity. The interaction starts off with the robot introducing itself and asking about the child's name and favourite food. This should make the child more comfortable and familiar with the robot and communication procedure. Then, the robot explains the goal of the conversation, i.e. to discuss and prioritize different values, and then the functional part of the interaction starts. The robot moves and maintains a standing position during this introduction. For the main part of the interaction that follows, the robot sits down and remains still, in order to not distract from the conversational content. In the main part, the robot proposes $4 - 8$ different everyday, school-related scenarios to the children and asks them to choose between two possible ways to react to the proposed scenario/dilemma. An example scenario is that there is a new child at school and they enter the classroom for the first time. One option to react is to offer the empty seat next to you to make them feel

[1] https://robotwise.nl/

comfortable and get to know them (benevolence). The other option is to remain passive and wait and see what happens (conformity). Between the child and the robot is a screen (as displayed in Figure 3.1) on which images appear that serve as a visual aid to the robot's stories and the decisions that the child has to make. A picture of the setting can be found in Figure 3.5. The child verbally expresses which option they would choose in the proposed situation. The preference of one option over the other is stored in the robot's memory and provides information about the child's inner values. Subsequently, the child is asked questions about their reasoning behind this decision. Specifically, the robot asks *why* the child chose that option; the *difficulty level* of reaching a decision on a scale of $1-5$; and whether the child has had *previous encounters* of that type before. After completion of all these questions, the robot moves on to and introduces the next scenario to be discussed, or concludes the interaction and wishes the child goodbye.

3.4.2. Post-interaction interview

Immediately following the interaction with the robot, a moderator guides the child from the cHRI room to a separate interview room. The moderator does not discuss any experiment content with them so as to not disturb the natural memory process. During the interview, the interviewer sits behind a table and starts audio recording the conversation when the child takes place on the chair across the table. The child is then asked to reproduce as many scenarios/dilemmas that the robot proposed as possible, with as many details as possible. When the child is not able to produce any more information/recollections spontaneously, they are presented with printed copies of the visuals they had seen onscreen during the cHRI, as well as the visuals of scenarios that had not been presented to them. In total, they are presented with 16 visuals, of which they have seen either 4 or 8 during their own session. They can keep these visuals to look through them in order to refresh their memory. They are then asked to point out the scenarios that they remember (that they have not mentioned already) and to reproduce them with as many details as possible. This time, when they run out of things to say, the interview is over and as such the session for the child is over.

3.4.3. Sessions

According to the initial plan, all participants would discuss all 16 scenarios, through a predetermined schedule in which each participant would have 1 session per day (out of 2 days), in which 8 scenarios were discussed. The first 3 participants were presented with 8 scenarios as planned. However, in the interest of time, the cHRIs were then re-designed on-the-fly to discuss only 4 scenarios. This strategy proved to allow for more participants to contribute and thus for more diversity. At the end of the second day, both the moderators and children were more accustomed to the cHRI procedure, which made the interactions more efficient. Therefore, the last 2 participants on day 2 were also taking part in sessions with 8 scenarios each.

3.5. Measurements

As stated in section 3.1, there are two variables to be measured in this study, the eye gaze and the memorability of different moments. In turn, when a moment is considered memorable, it is sub-divided into the relevant class that describes the child's reason for remembering this moment, as well as the quality of the provided memory for the moment. The technical process of acquiring and processing these variables is described in this section.

3.5.1. Eye gaze

Gaze extraction As stated in chapter 1, the design of an eye gaze tracking algorithm is not within the scope of this project and this process is being taken care of by an algorithm designed by IDIAP. The raw camera footage of the children's faces, recorded from the cHRI, is fed through this algorithm. The output contains the following information for each frame of each video: (1) if the participant looks at the robot; (2) if the participant looks at the screen; (3) if the participant looks at any other point in space

Handpicked dataset During the gaze extraction process it was found that many cHRIs were interrupted once or multiple times by the moderator because of the robot malfunctioning. Because of this, the interactions are not purely between the child and the robot, but also with the moderator. This creates diverging gaze patterns that are not accurately picked up by the gaze extraction algorithm, since

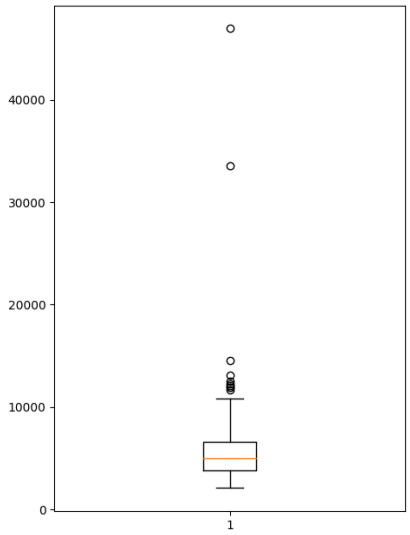
the moderator is not included as a visual target. In fact, the moderator was positioned behind the child, so some children were frequently turning their back to the camera for recording gaze. In addition, some children are more energetic or restless than others, causing them to move a lot. The gaze extraction algorithm is not perfectly attuned to this type of behaviour and thus struggles to keep up with rapid changes in head position. For this reason, it was decided to make an additional sub-selection of the data in which there were no or very few interruptions by the moderator and the child stays relatively still. This dataset is accumulated through manual visual analysis of the recordings by the researcher, where they made note of these well-executed cHRIs with well-adjusted participants. As a result, certain interactions are filtered out. So, a total of 14 children participated in this version, compared to 30 normally. 8 of the participants are male and 6 are female, almost the same ratio as in the full dataset ($6/8$ vs $12/18$). Their age ranges from $7-11$ years old, with a mean of 8.5 years ($SD = 1.2y$), the same as in the full dataset. The fact that the demographics from the handpicked dataset are similar to the ones in the full dataset suggests that the data quality is not dependent on the subjects but rather on hardware/technological issues. The dataset obtained through this process is called 'handpicked dataset' and contains 87 samples; 59 memorized and 28 not memorized, a ratio of 0.68 memorized samples. In the full dataset, the ratio is 0.61. The difference is small and could be due to the fact that interactions with many technical issues were filtered out. These had interruptions in the scenario discussions and could therefore be less memorable. A full description and class division of the final datasets can be found at the end of this chapter.

Identification of 'moments' In order to provide structure to the gaze data, the data has to be divided into distinguishable 'moments' that are each considered one sample during both labelling and classification. It has been predefined that the moments that will be taken as samples are those that correspond to a single scenario discussion. The starting point of a scenario discussion is defined as the moment that the robot introduces a new value-based dilemma, which corresponds to the moment in time that the image on the screen is changed/updated. The end point of a scenario discussion occurs when the robot has verbally obtained all the desired information from the child regarding that scenario, and is defined as the moment that the robot switches to the next scenario/image on screen. The NAO robot's log files can be used to retrace the points in time (or frames) during which the robot switches to a new visual. Using these checkpoints, the eye gaze dataset is split up into 208 samples. The time chunks last between $2092-46987$ frames, i.e. $70-1566$ seconds with 30fps, with a 0.25 quantile of 3771 frames/126 seconds, 0.75 quantile of 6623 frames/221 seconds and an average length of 5802 frames/193 seconds and standard deviation of 4254 frames/141 seconds. As can be seen in Figure 3.6a, the time chunk durations are not distributed evenly and there are a lot of outliers on the upper side. The outliers are defined as being 1.5 times the interquartile range (0.75 quantile -0.25 quantile) greater than the 0.75 quantile. These outliers appear for two reasons. Firstly, there were some issues with the cHRI, especially on the first day, that caused for some restarts and re-dos that caused the start and end time of the scenario discussions to not be translated well to the log files. Another reason is that in some cases the participants were kept in the cHRI room longer than necessary, to enable smooth logistic organisation. This meant that, at times, the activity was left open and the scenario is thus falsely recorded for too long. To ensure the quality of the data, the outliers displayed in Figure 3.6a are discarded from the dataset. The filtered dataset contains 196 samples and has a sample length distribution as displayed in Figure 3.6b. This distribution has a 0.25 quantile of 3742 frames/125 seconds, a 0.75 quantile of 6322 frames/211 seconds and a mean value of 5177 frames/173 seconds and standard deviation of 2027 frames/68 seconds.

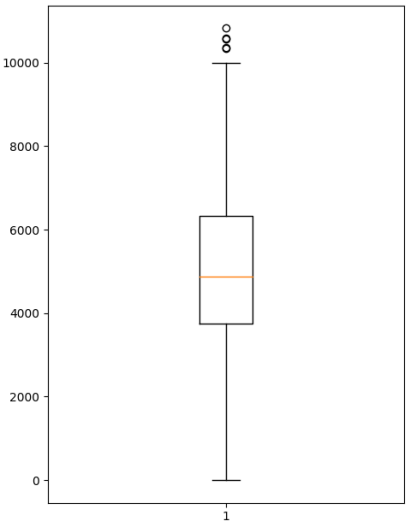
Sample length Despite recent progress in time series classification algorithms, most algorithms do not support time series of variable length, especially in the multivariate case. In order to use state-of-the-art technology, it was decided to alter the time chunks in the dataset to all be of the same length. It is decided to determine a cutoff value, cut and do the following for $x \in TimeChunks$:

$$len(x) < cut \Rightarrow discard(x)$$
$$len(x) \geq cut \Rightarrow return(x[0 : cut])$$

The scenarios, which make up the possible memorable moments, were originally designed to last for 120 seconds/3600 frames. The time chunks that last the shortest are the ones in which the cHRI was



(a) A boxplot of the different durations (in frames) of scenario discussions in the cHRI.



(b) A boxplot of the different durations (in frames) of scenario discussions in the cHRI after removing outliers.

Figure 3.6: The lengths of all the time chunks in the gaze dataset.

executed the most easily and smoothly and are thus valuable in terms of quality. In addition, during each scenario discussion, the children are introduced to the scenario and making their decision in the first part of the discussion. Hereafter, they talk about their decision making process. So, arguably, the first part of each time chunk is the most interesting or useful in terms of eye gaze patterns as well. Therefore, it is decided that cut should be of minimal value, while still retaining enough information. It is decided to use the 0.25 quantile value of the boxplot in Figure 3.6b as cut, i.e. 3742 frames, since it allows to retain 75% of the data while still being sufficiently long to withhold the majority or most relevant part of the scenario discussions. Applying this cutoff strategy to the dataset results in a final dataset of 149 samples, each of length 3742 frames/125 seconds.

3.5.2. Memorability and sub-labels

The information regarding whether a moment, i.e. a scenario discussion during the cHRI, is relevant or not is acquired in the post-interaction interview. In addition, the post-interaction interview contains information regarding the quality of the reproduced memory and the reason for remembering. As described in section 3.1, these interviews are held verbally and are semi-structured. The process that is required to extract (structural) meaning from this data will be described in this section.

Transcription Following the data collection there is a total of ± 9 hours of $.wav$ audio files. For easier analysis and annotation, these audio files are transcribed to text. Given the magnitude of the dataset, local automatic speech to text recognition (ASR) is used for the initial transcription. Following the automatic transcription with Whisper, the accuracy and correspondence of all generated files was verified and manually corrected where needed by the researcher. In addition, the utterances were manually annotated with who is responsible for saying it: either the interviewer or the participant. About half of the audio files collected on day 1, 23% of the total dataset, resulted in unusable transcriptions by Whisper because the microphone was ill-adjusted and the utterances, especially those of the participants, are hard to understand. Some audio manipulation, that is, maximum amplification and a noise reduction effect from Audacity's default effect library, was necessary to manually transcribe these files. Finally, in order to ease the analysis and annotation processes, the data was combined into a single, column-structured storage file that has each utterance on a new line, along with the person responsible for producing the utterance, the participant ID and the day of the recording. This dataset will be referred to as *interview_text*.

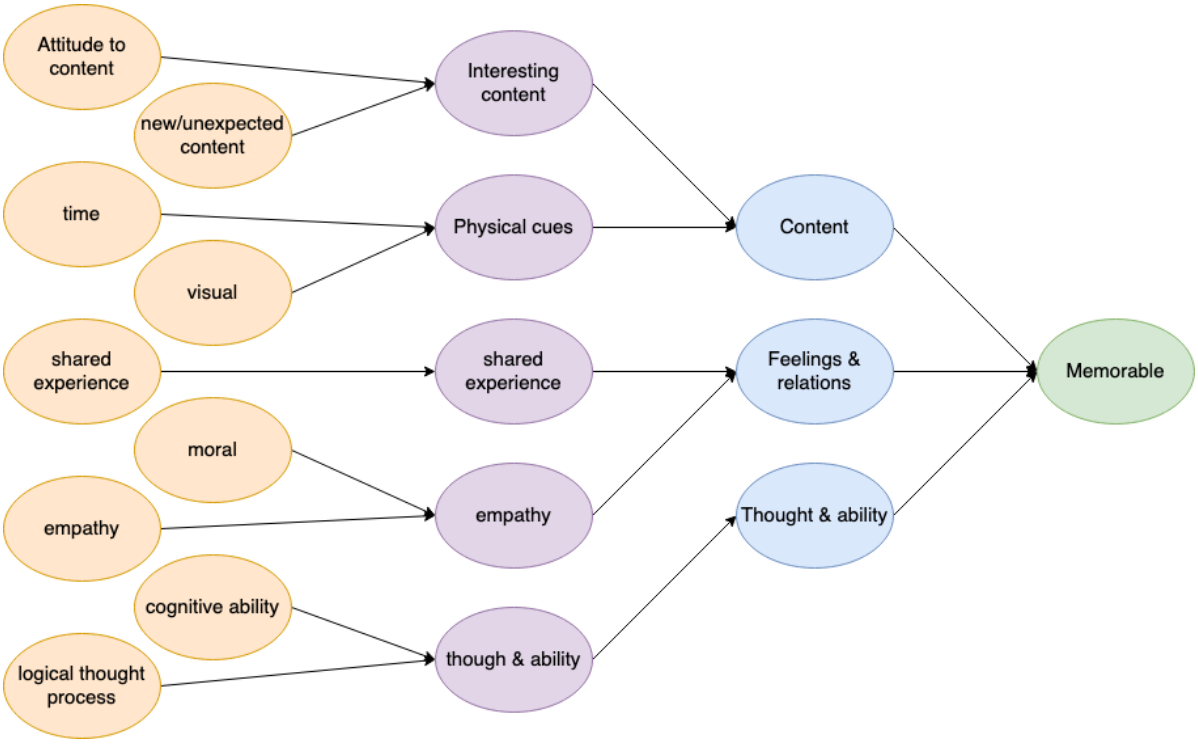


Figure 3.7: The three top levels of the thematic hierarchy that capture the children's reasons for remembering moments of the cHRI.

Memorability Given *interview_text*, the moments referred to by the children should be labelled as references of memorable moments. To be more precise, a child 'referring' to a moment could be any mention, both vague ('the one with the sandwich') and more specific ('I had to choose between eating my own disgusting lunch and asking a classmate for his pancakes.'), of one of the scenarios. As such, each line in *interview_text* in which a participant refers to a specific scenario is labelled with the ID of this scenario, which is a number in the range $[1-16]$ for each of the 16 scenarios. Following this process, the identification of memorable moments is complete. This dataset will be referred to as the 'binary dataset'.

Thematic analysis Based on the reflexive thematic analysis method described in section 3.1, the full dataset is read and analyzed multiple times over by the researcher, each time refining and altering the suggested themes that emerge from *interview_text*. This process has been schematically represented in Figure 3.7. The result of the thematic analysis process is a hierarchy of different themes, or rather labels, which are listed in Table 3.2. A discussion on how these labels relate to existing literature in the field can be found in chapter 5. In addition, some examples and excerpts of the interview transcriptions can be found in Appendix B. Applying the final set of labels to *interview_text*, thus flagging and labelling the relevant moments, results in a dataset called the 'reason dataset'.

Memory quality In addition to labelling each remembered moment with one of the labels obtained through thematic analysis, they are also labelled according to the 'quality' of the memory (if remembered). The quality of the memory ranges from *low* (1) to *medium* (2) to *high* (3) and is defined by the rules stated in Table 3.1. These rules were drawn up by the researcher, based on common behavioural/conversational patterns that were detected in *interview_text* during analysis of the data and the thematic analysis. These rules are then simply applied to the full dataset. The resulting labeled dataset is called the 'quality dataset'.

3.5.3. Machine learning adaptation

Following the data preparation processes described above, the data content is complete. The next step is to apply the machine learning models chosen in section 2.8, i.e. ROCKET, DrCIF, STC, TDE

labels	sublabels	description
content	interesting_content	the child is surprised by the presented content and/or has an opinion about it
	physical_cues	the child is triggered by visual or temporal content to remember
feelings_&_relations	shared_experience	the child has experienced or felt something similar as discussed
	empathy	the child can imagine the feelings of others in the scenario discussed
thought_&_ability	thought_&_ability	the child has (exceptional) ease or difficulty processing the content and/or making decisions
no_reason	no_reason	the child is not able to identify a reason for remembering a moment, nor do they provide motivated or coloured descriptions of their memory

Table 3.2: The final labels for the children's self-identified reasons for remembering parts of their interaction with the robot, as obtained through reflexive thematic analysis and the heuristics from Table B.1.

and HIVE-COTE v2, to the datasets (binary dataset, handpicked binary dataset, reason dataset and quality dataset) and train them to achieve maximum performance.

Training and validation The final datasets contain labelled samples of time series chunks of length 3742 frames, with a class division as shown in Figure 3.8, Figure 3.9, Figure 3.10 and Figure 3.11. As is apparent in these tables, the datasets are slightly unbalanced. To resolve this, the data is undersampled, during training only, to match the lowest number of samples in any of the classes, using random selection without replacement. The models are evaluated using 6-fold cross validation.

Model hyperparameters As described in section 2.8, ROCKET has a relatively simple architecture and is based on the use of convolutional kernels. Therefore, the only hyperparameter to be set for this model is the number of kernels. The library's default value is set at 10000, but model finetuning found better results for 1000 kernels.

HIVE-COTE v2 is an ensemble that is constituted of the following members: STC, DrCIF, Arsenal, and TDE. It takes as input parameters all the hyperparameters of the underlying models, which means that we find the hyperparameters for HIVE-COTE v2 at the same time that we find those for the other models, most of which are also run separately in this study. The hyperparameters that were found for the different models can be found in Table 3.3. Since DrCIF is an ensemble, it allows the user to set the number of estimators within the ensemble as a parameters. Additionally, it allows to define the number of intervals to extract per sample and the number of summary statistic attributes to subsample per tree. For STC, the hyperparameter is the number of shapelets to be extracted, that should define class correspondence. Arsenal is an ensemble of Rocket classifiers, and so it allows to set both the number of kernels per Rocket classifier (which will be kept the same as for the single Rocket classifier as defined above), as well as the number of estimators to include in the ensemble. TDE is an ensemble as well, but it automatically optimizes the number of estimators. Nevertheless, one can set a maximum ensemble size to limit the resource consumption. TDE also allows to finetune the number of parameters/summary statistics to extract from the samples.

HC2 member	parameter	value
DrCIF	n_estimators	30
	n_intervals	5
	att_subsample_size	10
STC	n_shapelet_samples	10000
Arsenal	num_kernels	1000
	n_estimators	30
TDE	n_parameter_samples	250
	max_ensemble_size	100

Table 3.3: The parameters for the HIVE-COTE v2 model, listed per member of the ensemble.

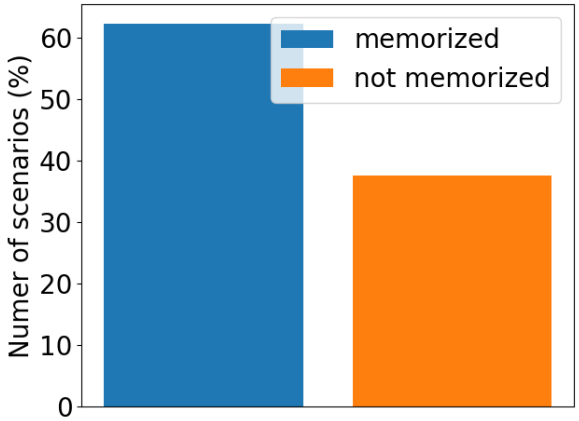


Figure 3.8: The label distribution in the binary dataset.

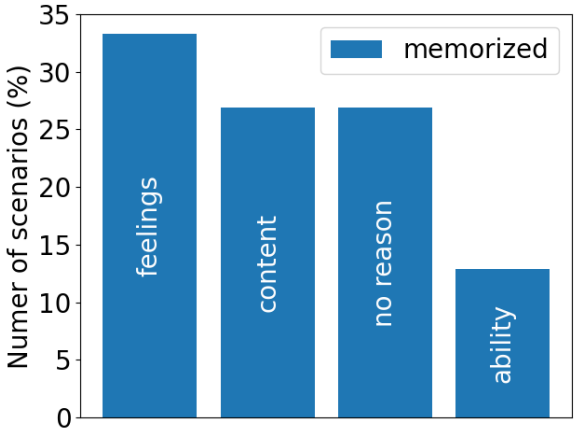


Figure 3.9: The label distribution, within all memorized moments, in the reason dataset.

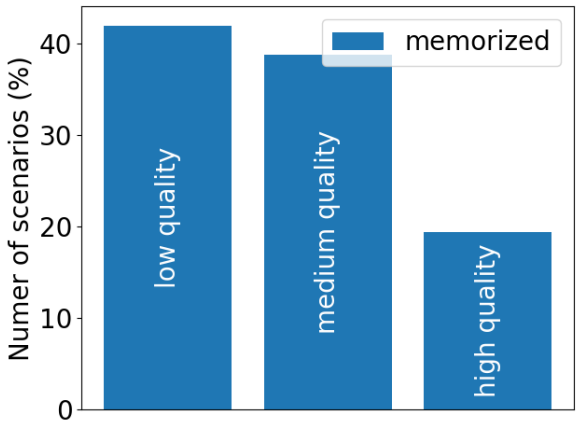


Figure 3.10: The label distribution, within all memorized moments, in the quality dataset.

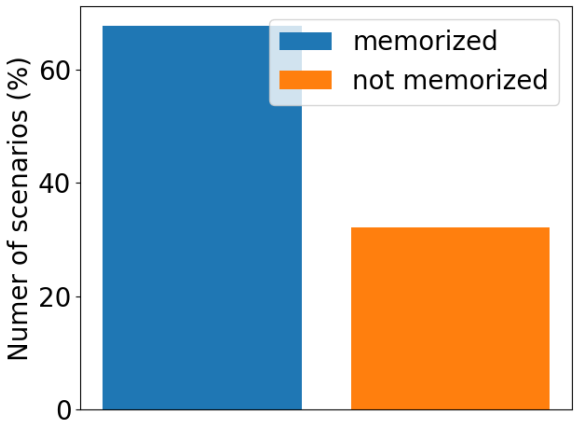


Figure 3.11: The label distribution in the handpicked dataset.

3.5.4. Data analyses adaptation

Sample length In order to successfully analyze the data characteristics, the data needs to be arranged such that it is suitable for statistical analyses. First off, whereas intelligent time series classifiers are equipped to handle long time series, statistical tests are not designed to handle these temporal characteristics. In order to perform statistical tests on time series data, the data has to be aggregated, in batches of size l, over the time series. If l is too large it only captures very general trends and if l is too small it does not capture meaning, or any of the temporal features. In literature, l ranges between $5-30$ seconds. For this study, all the general trends over the dataset, i.e. all visual target distributions, are computed with $l = 30s$ and analyses over time (aggregated over all the moments) are computed with $l = 5s$.

Gaze alternations Saccades are traditionally measured by taking each small directional change in eye gaze into account This information is not available in a dataset that records only the attention to visual targets. Nevertheless, there is likely still information to be gained from the gaze switching between the visual targets. In order to access this information, a loop over each 30 seconds records a counter of the occurrence of different tri-grams. A tri-gram, in this context, is defined as three consecutive gaze targets, where two neighbouring targets can not be the same, e.g. $robot - screen\_left - other$, $screen\_left - screen\_right - screen\_left$ and $other - robot - other$, but not: $robot - robot - other$. The relevant tri-grams for the analyses are $screen\_left - screen\_right - screen\_left$ and $screen\_right - screen\_left - screen\_right$, which can be summed together to form the $within\_screen$ gaze alternations. The occurrence of these $within\_screen$ gaze alternations can be taken relative to the total number of gaze alternations in the time frame ($30s$), as well as relative to the total time spent gazing at the screen within these $30s$.

Fixation length Another common metric for gaze pattern analyses is the duration of eye gaze fixation in a single point. The exact coordinate point of gaze direction is not available, however, there is likely still information to be gained from measuring the fixation duration on a single visual target. In order to access this information, a loop over each 30 seconds records the number of frames during which the gaze target remains unchanged, for each subsequent gaze target. Then, the average fixation length for each visual target during that time frame ($30s$) is computed.

3.6. Data analyses

Based on the findings on gaze pattern heuristics in chapter 2, it will be researched which of these heuristics can be transferred to memorable moments detection during cHRI and what other heuristics in this field can be found. This will be done according to statistical tests on the collected dataset. After conducting Shapiro-Wilk tests for normality, in which the null hypothesis that the data follows a normal distribution was rejected ($p < .001$), it was decided that the following statistical tests will be used in the analyses: ANOSIM and Mann-Whitney U. These are the non-parametric verisons of MANOVA and the independent t-test, respectively. ANOSIM will be used to test for main effects of the different labels over multiple measurements. Mann-Whitney U tests will be used to test for significant difference between the labels for one measure. It will be researched if the data supports the hypotheses described in the following subsections.

3.6.1. Visual target distribution

As a test for validity and to investigate the magnitude of the memorable versus not memorable effect, it will be researched if there are significant differences between the global distribution of the time spent looking at each visual target. This should provide an insight into if the samples categorized as memorable exhibit different gaze patterns compared to the not memorable case on a large scale, which should encourage further and deeper investigation into the specifics of these differences.

Hypothesis 1: Children who are likely to remember a moment during the cHRI exhibit different gaze-time distributions over the different visual targets, compared to children who are not likely to remember a moment.

In order to test the validity of this hypothesis, the gaze time distribution is aggregated per 30 seconds and stored as the relative time spent looking at each visual target during these 30 seconds. Consecutively, it will be researched if there is an overall difference between the memorable and the not

memorable case with an ANOSIM test and it will be researched if there are differences between the two cases per visual target ('screen', 'robot' and 'other') using Mann-Whitney U tests. In addition, the same tests will be executed to compare the memorable samples with the non-memorable samples, except the 'low quality' memorable samples are removed from the dataset. This is due to the expected noise in the ground truth dataset that occurs because children's self-identification of memorable moments is questionable when they refer to the moment with little to no information.

Given the emphasis on cHRI in this study, it is especially relevant to investigate whether children who are likely to remember a moment are more likely to gaze at the robot, compared to children who are not likely to remember a moment. As stated in chapter 2, perhaps a stronger bond between the child and the robot, indicated by more mutual gaze, would lead to a higher likeliness of remembering.

3.6.2. Visual target distribution over time

At the start of each scenario discussion, the robot introduces the scenario and provides the two available options that the child must choose between. The mean duration of this introductory monologue is 35.2s ($SD = 5,9$s). During this introduction, eye gaze patterns are expected to be different from the ones after the introduction for multiple reasons. First of all, the nature of the interaction at this moment is more one-sided, as compared to the rest of the interaction that consists of a shorter question-answer structure. In addition, the child is presented with a lot of new information: the visual on screen changes and the robot talks through the entire scenario and explains the choice to be made. Finally, beyond this point of the interaction, the behaviour of both the robot and the child relies on many factors that may not be directly related to the interaction content. For example, the robot may face technical issues or have difficulties understanding the child, which could, in turn, lead to the child getting distracted or annoyed. For these reasons, the following hypothesis will be included in the research:

Hypothesis 2: Gaze patterns at the start of a scenario, i.e. when the scenario topic/content is introduced (≤ 35s), are more indicative of the scenario being memorable or not than those later in the scenario discussion.

In order to test the validity of this hypothesis, the gaze time distribution is aggregated per 5 seconds and stored as the relative time spent looking at each visual target during this time. It will be researched how the gaze distribution over the different visual targets differs between the memorized samples and not memorized samples per time step and how these differences change or deteriorate over time.

3.6.3. Gaze alternations

A central aspect of the scenario discussions in the cHRI is that the child makes a decision about a fictitious but realistic scenario, based on their morals. It could be that children who are indecisive regarding their decision are less prone to remember that moment, due to their brain activity being elsewhere than memorization. Therefore, the following hypothesis was drawn up:

Hypothesis 3: Children are less likely to remember scenarios if they are indecisive regarding the choice to be made during these scenarios, compared to children who appear decisive.

In order to find out whether this hypothesis holds, it needs to be determined how to evaluate decisiveness. For this problem setting, it will be assumed that the number of gaze alternations between the two options as displayed on screen correlates with decisiveness. For memorable moments, it is expected that the number of within-screen gaze alternations, relative to the total number of alternations aggregated per 30 seconds, is significantly lower than during non-memorable moments. This will be tested with a Mann-Whitney U test. In addition, the duration of gaze fixations might also be positively correlated with indecisiveness, as longer fixations are indicative of deeper inner thoughts. To investigate this, it will be tested with a Mann-Whitney U test whether average fixations durations are significantly shorter in memorable moments compared to those in non-memorable moments.

3.6.4. Ease of understanding

As mentioned before, it can be expected that children are less likely to remember moments when their brains are preoccupied with other processes in those moments. This preoccupation may not only be caused by the difficulty of the choice, but also by the difficulty of the entire scenario. Generally, participants who were not able to reproduce/remember any of the scenarios from the cHRI indicated that they found the questions or content to be complicated. On the other hand, a significant portion of the children who remembered moments indicated to remember that moments due to it being 'easy'. Opposite to the hypothesis regarding indecisiveness, it is expected that children who remember moments

due to them being easy are spending less time looking at the screen and more time looking at ‘other’ and ‘robot’. The following hypothesis was established:

Hypothesis 4: Children who self-identified their reason for remembering a scenario to be its (cognitive) ease spend more time looking at ‘other’ and ‘robot’ and less time looking at ‘screen’.

To investigate the validity of this hypothesis, Mann-Whitney U tests will be applied to the relative time spent looking at ’robot’ and ’other’, aggregated per 30 seconds, for reason-label ’ability’ compared to the other reasons.

3.6.5. Memory quality

The memories of the children are also given a memory quality label. As per its definition, moments that are considered memorable with a high quality memory are described elaborately, including descriptions of the items or characters on screen (see Table 3.1). It is expected that children who can reproduce a lot of these (visual) details are spending more time looking at the screen. To investigate this, the following hypothesis was proposed:

Hypothesis 5: Children who can reproduce ‘high quality’ recollections of a scenario spend more time looking at the screen during these scenarios.

The validity of the hypothesis will be tested with a Mann Whitney U test that will compare the time spent looking at ‘screen’ between the high quality memorable moments and the low and medium quality memorable moments. The time spent looking at ‘screen’ is relative to the time spent looking at other visual targets, aggregated per 30s. Similarly, a Mann Whitney U test will be conducted that will compare the average fixation durations on visual target ‘screen’ between the high quality memorable moments and the low and medium quality memorable moments.

3.6.6. Feelings versus content

As mentioned in hypothesis 5, it is expected that children who focus on the content of the scenario discussion spend more time looking at the screen. Conversely, children who focus more on the social aspect might be looking more at the robot, as a social connection, or more at ‘other’, to process their feelings or emotions. Based on this theory, the following hypothesis is proposed:

Hypothesis 6: Children who indicate that they remember a scenario because of how it made them feel are looking more at ’robot’ and ’other’ compared to children who indicate to remember a scenario because of the content, who, in turn, are looking more at ‘screen’.

The validity of the hypothesis will be tested with a Mann Whitney U test that will compare the time spent looking at each visual target (‘screen’,‘robot’,‘other’) between the moments memorable due to feelings and those memorable due to the content. The time spent looking at each target is relative to the time spent looking at other visual targets and is aggregated over each 30s. A Mann-Whitney U test will also be executed over the average fixation times for these two classes.

4

Results

Following the method described in chapter 3, results are obtained in the form of trained classifiers and statistics. With these results, the objective of this study should be fulfilled. Namely, it should become clear how and to what extent eye gaze tracking during cHRIs can be exploited to identify moments that the child remembers from this interaction. The results should shed light on what a child remembers, why they remember this and to what extent they remember this moment, using the different datasets described in chapter 3. The hypotheses established in chapter 2 will guide the analysis of the results.

4.1. Binary dataset

The binary dataset consists of 149 moments during the cHRIs that are classified as either memorized ($n = 93$) or not memorized ($n = 56$) (see Figure 3.8). This means there are a total of 2 classes in this dataset. With this dataset, it was researched what patterns of meaning can be found between the different eye gaze patterns for memorable moments and not memorable moments. Then it is verified if the machine learning models can pick up on these patterns automatically.

4.1.1. Visual target distribution

Hypothesis 1: Children who are likely to remember a moment during the cHRI exhibit different gaze-time distributions over the different visual targets, compared to children who are not likely to remember a moment.

The overall visual target distribution is visualized in Figure 4.1. An ANOSIM test was carried out over this data that found no significant difference overall between the memorable and not memorable case ($p = .195$). Nevertheless, when only considering the memorable moments that are considered memories of medium and high quality, an ANOSIM test indicates an overall significant difference ($p = .0.005$) between the two groups. The distribution of the not memorized case and the medium quality memorized case is visualized in Figure 4.2. In addition, Mann-Whitney U tests on each individual visual target, for the memorable and non memorable case, indicated a significant difference between these cases for both visual target 'screen' ($MWU = 112875.5, n_1 = 370, n_2 = 559, p = 0.018$, Effect size by rank-biserial correlation: negligible) and 'other' $MWU = 95218.0, n_1 = 370, n_2 = 559, p = 0.040$, Effect size by rank-biserial correlation: negligible). As expected, the difference between these cases when only including the medium/high quality memory samples is also significant for both 'screen' ($MWU = 55081.0, n_1 = 370, n_2 = 239, p < 0.001$, Effect size by rank-biserial correlation: small) and 'other' ($MWU = 35025.0, n_1 = 370, n_2 = 239, p < 0.001$, Effect size by rank-biserial correlation: small). In addition, the time spent looking at the robot is significantly larger in the medium/high quality group compared to the not memorized group ($MWU = 37100.5, n_1 = 370, n_2 = 239, p < 0.001$, Effect size by rank-biserial correlation: small). It should also be noted that for the latter experiment the effect size was shown to be bigger in the medium quality case for all visual targets and thus the actual difference between the two cases is larger when considering only the medium/high quality label.

Summary Globally, there is no different gaze-time distribution over the visual targets between the memorized moments and the not memorized moments. However, when applying a stricter bound

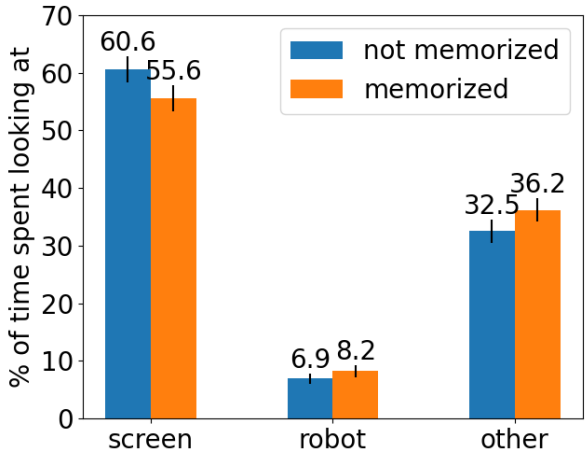


Figure 4.1: The relative time spent looking at the 3 different visual targets, aggregated per 30 seconds and grouped by the labels not memorized and memorized. The mean and 95% confidence interval are displayed in black.

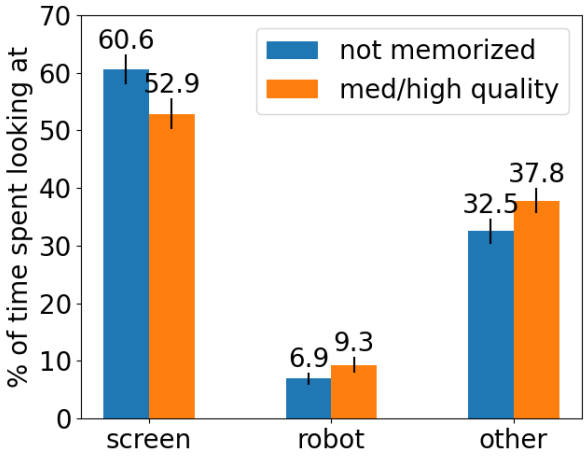


Figure 4.2: The relative time spent looking at the 3 different visual targets, aggregated per 30 seconds and grouped by the labels not memorized and 'medium/high quality' memorized. The mean and 95% confidence interval are displayed in black.

considering moments memorable or not, i.e. when using only the medium and high quality label, there is a significant main effect between the two cases. During medium/high quality memorable moments, children look significantly more at 'robot' and 'other' and significantly less at 'screen'. These differences in gaze behaviour, apparent on a large scale, provide enough reason for deeper investigation.

4.1.2. Visual target distribution over time

Hypothesis 2: Gaze patterns at the start of a scenario, i.e. when the scenario topic/content is introduced (≤ 35s), are more indicative of the scenario being memorable or not than those later in the scenario discussion.

The plot in Figure 4.3 shows the average time spent gazing at each visual target throughout a scenario discussion, for both memorable (M) and not memorable (NM) moments. From this plot, it appears as though, indeed, the gaze target distribution at the start of the scenario discussions are more distinctive between the two classes than towards the end of the scenario. The Mann-Whitney U tests conducted over each 5-second fragment confirm this vision and provide significant ($p < .05$) differences for a majority of the intervals within the first 35 seconds. The results of these tests are shown as the shaded areas in Figure 4.3, where it can be seen that the significant differences occur for both visual targets 'screen' and 'other' in the first 30 seconds. For visual target 'robot', however, there appears to be no significant difference between the memorized and not memorized samples in the first 35 seconds. The memorability distinction based on target 'robot' is, instead, more apparent in the middle and end of the scenario discussion.

4.1.3. Gaze alternations

Hypothesis 3: Children are less likely to remember scenarios if they are indecisive regarding the choice to be made during these scenarios, compared to children who appear decisive.

To test this hypothesis, the number of trigrams of gaze alternations of type $screen\_left - screen\_right - screen\_left$ or $screen\_right - screen\_left - screen\_screen - right$, relative to the total number of trigrams aggregated per 30 seconds were compared between the labels not memorized and memorized. In Figure 4.4 it can be observed that children who remember moments are less likely to alternate between the two options provided on screen. This difference is more prominent in the first 60 seconds of the scenario discussion. In addition, the average fixation durations on the screen as a whole are displayed in Figure 4.5. The Mann-Whitney U tests for the relative number of gaze alternations within the screen proved that children's gaze alternates significantly less ($MWU = 98335.0, n_1 = 344, n_2 = 526, p = 0.029$, Effect size by rank-biserial correlation: negligible) between the two options on the screen during memorable moments compared to non-memorable moments. The same test,

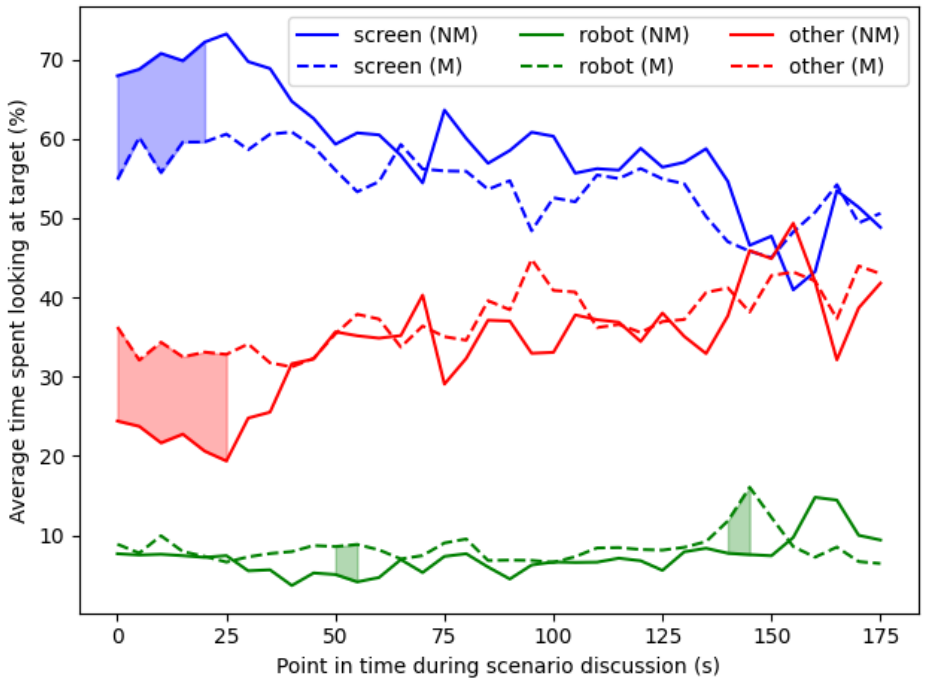


Figure 4.3: The average time spent looking at the 3 different visual targets throughout the duration of a scenario discussion, in steps of 5 seconds and grouped by the labels not memorized (NM) and memorized (M). The shaded areas are the intervals in which a Mann-Whitney U test found significant ($p < 0.05$) differences between the two groups (NM vs M) for each visual target.

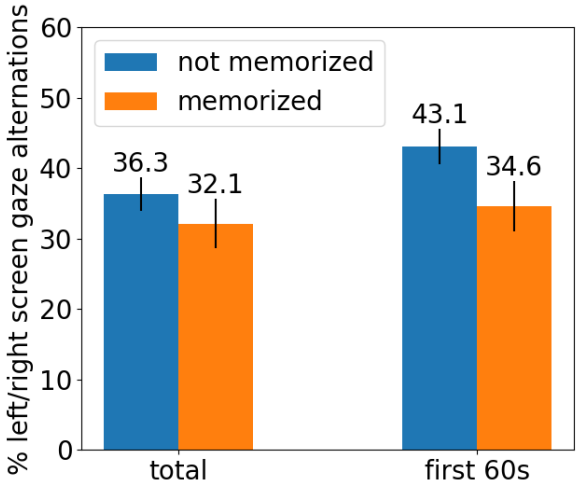


Figure 4.4: The number of within screen gaze alternations, relative to the total number of trigrams aggregated per 30 seconds and grouped by the labels not memorized and memorized. The statistic is shown over each entire scenario and over the first 60 seconds of each scenario. The mean and 95% confidence interval are displayed in black.

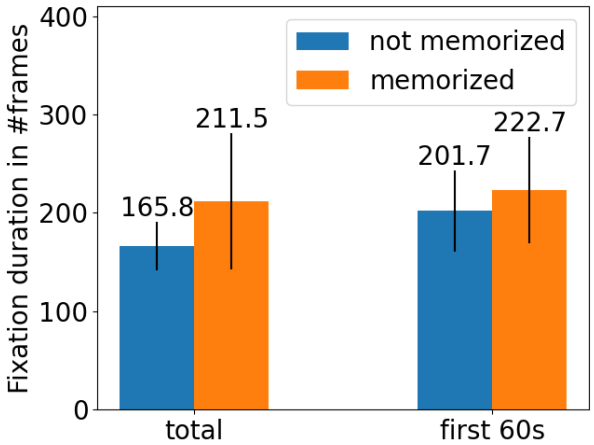


Figure 4.5: The average duration of the fixations on visual target 'screen', aggregated per scenario and grouped by the labels not memorized and memorized. The statistic is shown over each entire scenario and over the first 60 seconds of each scenario. The mean and 95% confidence interval are displayed in black.

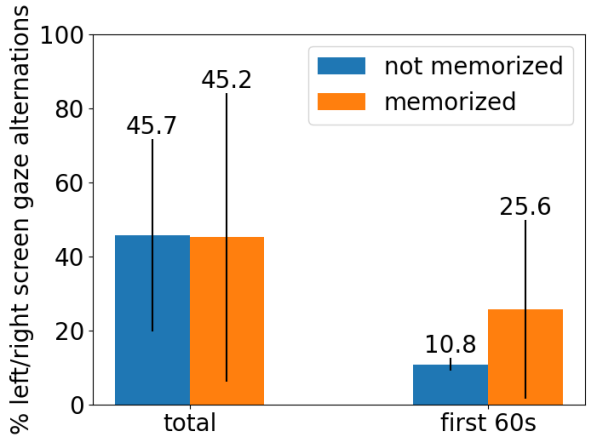


Figure 4.6: The number of within screen gaze alternations, relative to the time spent looking at the screen aggregated per 30 seconds and grouped by the labels not memorized and memorized. The statistic is shown over each entire scenario and over the first 60 seconds of each scenario. The mean and 95% confidence interval are displayed in black.

conducted on only the first 60 seconds of each scenario discussion, resulted in a significant difference as well, but with a larger effect size ($MWU = 17992.0, n_1 = 141, n_2 = 222, p = 0.016$, Effect size by rank-biserial correlation: small). The Mann-Whitney U tests for the fixation durations on screen during memorable moments showed no significant difference ($MWU = 4640.0, n_1 = 74, n_2 = 116, p = 0.347$) to the fixation durations on screen during non-memorable moments taken over the entire scenario, as well as taken only over the first 60 seconds ($MWU = 4552.0, n_1 = 72, n_2 = 113, p = 0.173$). When comparing the number of within screen gaze alternations between memorized and not memorized sampels, and taking these as a percentage over the total time spent looking at the screen, there is no significant effect. Figure 4.6 displays the distribution of this data, taken over all data and taken over the first 60 seconds. The Mann-Whitney U test for the total duration shows no significant effect ($MWU = 4603.5, n_1 = 74, n_2 = 111, p = 0.164$) and neither does the test on the first 60 seconds ($MWU = 842.0, n_1 = 29, n_2 = 47, p = 0.086$).

4.1.4. Machine learning model

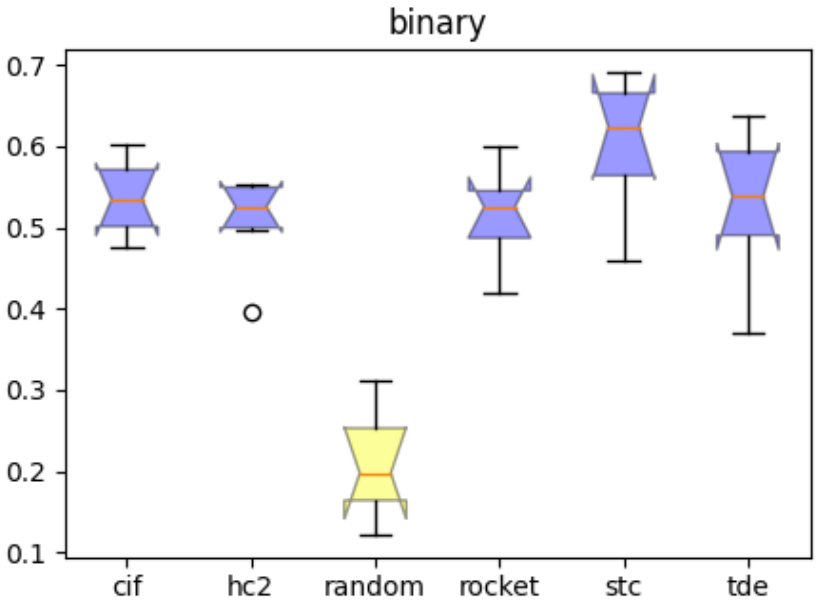


Figure 4.7: Boxplots of the performance of each model for the binary dataset. The models that performed significantly better than the dummy classifier (highlighted in yellow) are highlighted in blue.

In order to verify whether the patterns found in this dataset can be automatically detected, the different machine learning models, DrCIF, ROCKET, STC, TDE and HIVECOTE-v2 were trained on the binary dataset. The results of this are plotted in Figure 4.7. In order to meaningfully analyze the model's performance, they are compared with the performance of a random (dummy) classifier through independent t-tests.

There was a significant difference in the F1 score for problem 'Binary dataset' between DrCIF ($M = .209, SD = .07$) and Dummy ($M = .536, SD = .05$); $t(12) = 9.285, p < .001$. There was a significant difference in the F1 score for problem 'Binary dataset' between ROCKET ($M = .209, SD = .07$) and Dummy ($M = .516, SD = .062$); $t(12) = 8.007, p < .001$. There was a significant difference in the F1 score for problem 'Binary dataset' between STC ($M = .209, SD = .07$) and Dummy ($M = .603, SD = .087$); $t(12) = 8.633, p < .001$. There was a significant difference in the F1 score for problem 'Binary dataset' between TDE ($M = .209, SD = .07$) and Dummy ($M = .528, SD = .096$); $t(12) = 6.565, p < .001$. There was a significant difference in the F1 score for problem 'Binary dataset' between HIVE-COTE v2 ($M = .209, SD = .07$) and Dummy ($M = .508, SD = .06$); $t(12) = 7.935, p < .001$. In summary, for problem 'Binary dataset', every model performs significantly better than the random model.

4.1.5. First 60 seconds model

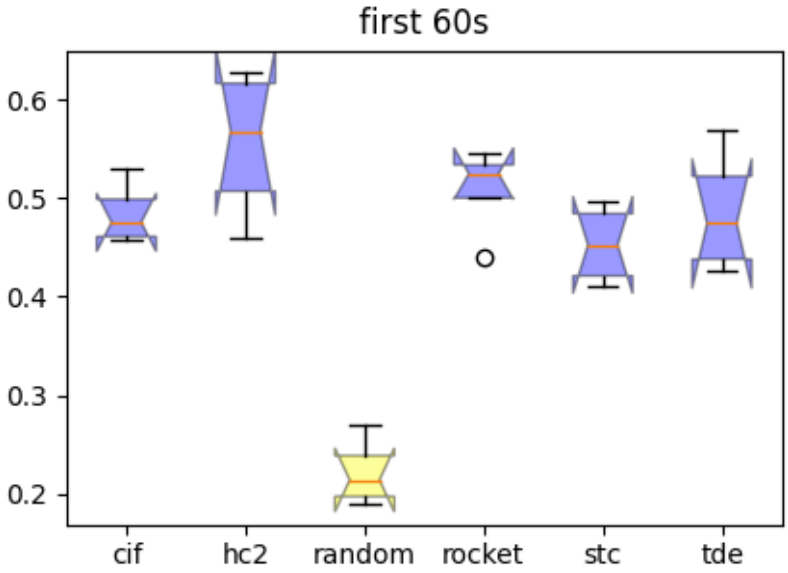


Figure 4.8: Boxplots of the performance of each model for the binary dataset, but only the first 60 seconds of each scenario discussion, i.e. 'moment'. The models that performed significantly better than the dummy classifier (highlighted in yellow) are highlighted in blue.

Based on the findings regarding the peculiarities for the moment in time during the scenario discussions, it was decided to additionally train the machine learning models on the binary dataset with only the first 60 seconds of each interaction in each sample. The results of this are plotted in Figure 4.8. In order to meaningfully analyze the model's performance, they are compared with the performance of a random (dummy) classifier through independent t-tests.

There is a significant difference in the F1 score for problem 'Binary dataset' between DrCIF ($M = .222, SD = .035$) and Dummy ($M = .484, SD = .033$); $t(8) = 10.891, p < .001$. There is a significant difference in the F1 score for problem 'Binary dataset' between ROCKET ($M = .222, SD = .035$) and Dummy ($M = .509, SD = .047$); $t(8) = 9.746, p < .001$. There is a significant difference in the F1 score for problem 'Binary dataset' between STC ($M = .222, SD = .035$) and Dummy ($M = .453, SD = .042$); $t(8) = 8.413, p < .001$. There is a significant difference in the F1 score for problem 'Binary dataset' between TDE ($M = .222, SD = .035$) and Dummy ($M = .486, SD = .065$); $t(8) = 7.119, p < .001$. There is a significant difference in the F1 score for problem 'Binary dataset' between HIVE-COTE v2 ($M = .222, SD = .035$) and Dummy ($M = .555, SD = .079$); $t(8) = 7.677, p < .001$. In summary, for problem 'Binary dataset', every model performs significantly better than the random model, also when only considering the first 60 seconds of each scenario discussion.

4.2. Handpicked dataset

The handpicked dataset consists of 87 moments during the cHRIs that are classified as being either memorized ($n = 59$) or not memorized ($n = 28$) (see Figure 3.11). This means there are a total of 2 classes in this dataset. This dataset is a subset of the binary dataset and was created to minimize the amount of technical or logistical errors or gaps in the data. For this reason, the machine learning models were run immediately on this dataset, to verify whether it would increase the performance.

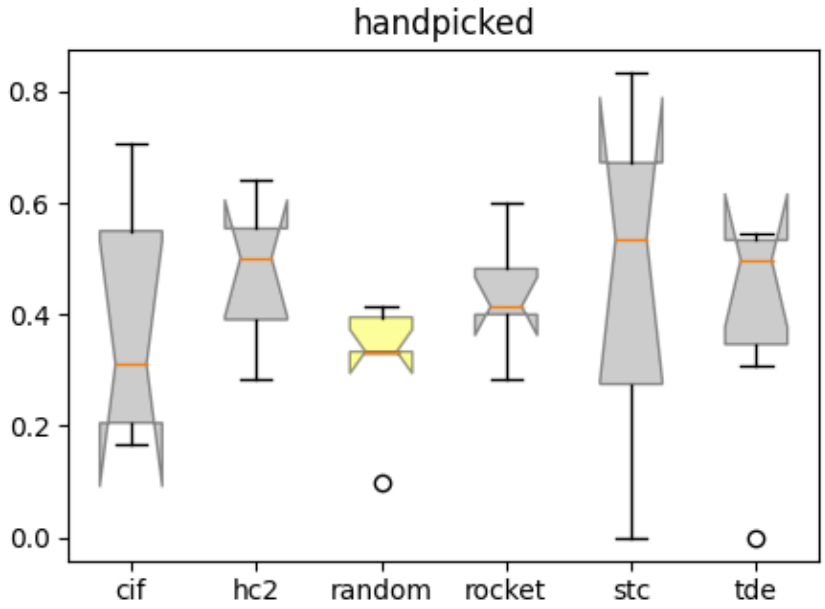


Figure 4.9: Boxplots of the performance of each model for the handpicked dataset. The models that performed significantly better than the dummy classifier (highlighted in yellow) are highlighted in blue.

The results of running DrCIF, ROCKET, STC, TDE and HIVECOTE-v2 on this dataset are plotted in Figure 4.9. In order to meaningfully analyze the model's performance, they are compared with the performance of a random (dummy) classifier through independent t-tests.

There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between DrCIF ($M = .322, SD = .116$) and Dummy ($M = .382, SD = .227$); $t(12) = .582, p = .573$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between ROCKET ($M = .322, SD = .116$) and Dummy ($M = .436, SD = .106$); $t(12) = 1.775, p = .106$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between STC ($M = .322, SD = .116$) and Dummy ($M = .468, SD = .314$); $t(12) = 1.073, p = .308$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between TDE ($M = .322, SD = .116$) and Dummy ($M = .397, SD = .214$); $t(12) = .755, p = .468$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between HIVE-COTE v2 ($M = .322, SD = .116$) and Dummy ($M = .476, SD = .133$); $t(12) = 2.135, p = .059$.

In summary, none of the models performed significantly better than the random classifier on the handpicked dataset. For this reason, it was decided to discard to dataset from further investigation.

4.3. Reason dataset

The reason dataset consists of 149 moments during the cHRIs that are classified as being either not memorized ($n = 56$) or memorized due to content ($n = 25$), feelings ($n = 31$), ability ($n = 12$), or no reason ($n = 25$) (see Figure 3.9). This means there are a total of 5 classes in this dataset. With this dataset, it was researched what patterns of meaning can be found between the different eye gaze patterns, within memorable moments, for different reasons for remembering, as identified in chapter 3. Then it is verified if the machine learning models can pick up on these patterns automatically.

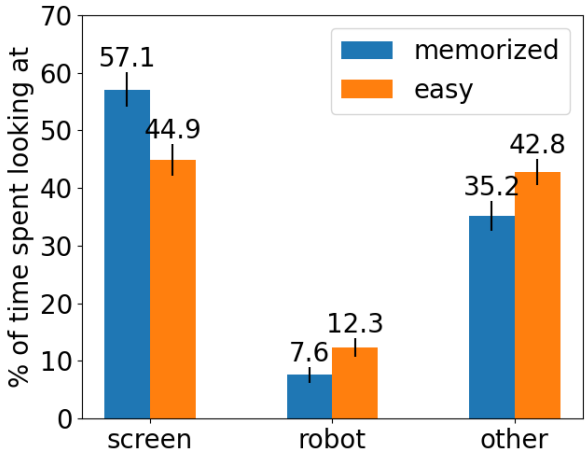


Figure 4.10: The relative time spent looking at the 3 different visual targets, aggregated per 30 seconds and grouped by the labels memorized (but not because of ease of understanding) and memorized because of ease of understanding. The mean and 95% confidence interval are displayed in black.

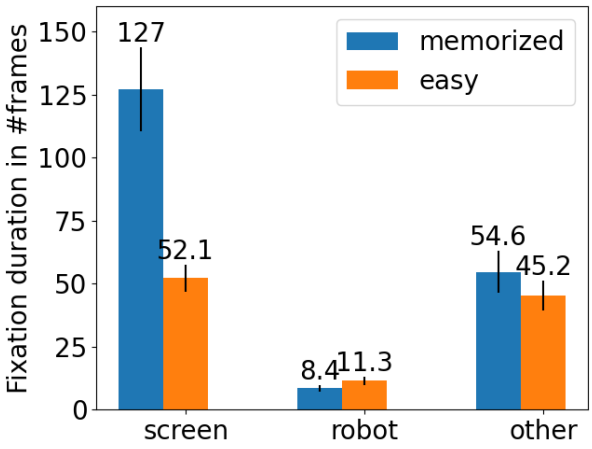


Figure 4.11: The average duration of the fixations on each visual target, aggregated per scenario and grouped by the labels memorized (but not because of ease of understanding) and memorized because of ease of understanding. The mean and 95% confidence interval are displayed in black.

4.3.1. Ease of understanding

Hypothesis 4: Children who self-identified their reason for remembering a scenario to be its (cognitive) ease spend more time looking at 'other' and at the robot and less time looking at the screen.

As can be seen in Figure 4.10, the data for the two groups follows a different distribution (ANOSIM, $p = .011$). This supports the hypothesis that, generally, children who remember a moment due to cognitive ease of the discussion content are more likely to gaze at both the robot ($MWU = 20659.5, n_1 = 72, n_2 = 487, p = 0.008$, Effect size by rank-biserial correlation: small) and 'other' ($MWU = 20499.0, n_1 = 72, n_2 = 487, p = 0.020$, Effect size by rank-biserial correlation: small) and less likely to gaze at the screen ($MWU = 14106.0, n_1 = 72, n_2 = 487, p = 0.007$, Effect size by rank-biserial correlation: small), compared to the other reasons for remembering. On the contrary, an ANOSIM test found no significant difference between the two groups for the average fixation duration ($p = .741$). Nevertheless, Figure 4.11 shows a particularly large difference in the fixation time on the screen, and so a Mann-Whitney U test was conducted to test if there is a significant difference between these groups. The result showed that the average fixation duration on screen is significantly smaller ($MWU = 10967.5, n_1 = 68, n_2 = 419, p = 0.002$, Effect size by rank-biserial correlation: small) for the cognitively easy moments compared to moments remembered due to other reasons.

4.3.2. Feelings versus content

Hypothesis 6: Children who indicate that they remember a scenario because of how it made them feel are looking more at 'robot' and 'other' compared to children who indicate to remember a scenario because of the content, who, in turn, are looking more at 'screen'.

The gaze time distribution over the visual targets for the reason labels 'content' and 'feelings' is displayed in Figure 4.12. An ANOSIM test found a significant main effect between the labels 'content' and 'feelings' ($p = .028$). A post-hoc Mann-Whitney U test found that there is no significant difference between the two groups for the time spent looking at the robot ($MWU = 14097.5, n_1 = 192, n_2 = 151, p = 0.628$). Nevertheless, children who remember a moment because of feelings spend significantly less time looking at the screen ($MWU = 12616.0, n_1 = 192, n_2 = 151, p = 0.039$, Effect size by rank-biserial correlation: small) and significantly more time looking at 'other' ($MWU = 17028.5, n_1 = 192, n_2 = 151, p = 0.005$, Effect size by rank-biserial correlation: small), compared to children who remember a moment because of the content.

The comparison of the fixation duration on each visual target, plotted in Figure 4.13 found no significant main effect ($p = .558$) between the two groups.

In Figure 4.14 the gaze target distribution for these two labels is plotted over time, and the shaded areas in the plot indicate the intervals in which there is a significant difference between the two groups. It can be seen that the differences between these groups occur in middle and end part of the scenario

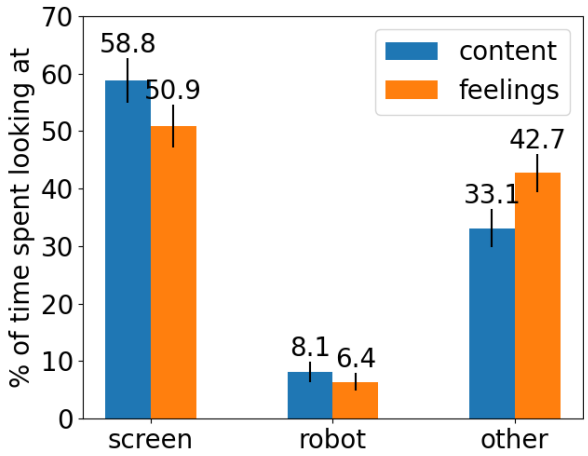


Figure 4.12: The relative time spent looking at the 3 different visual targets, aggregated per 30 seconds and grouped by the labels memorized because of content and memorized because of feelings. The mean and 95% confidence interval are displayed in black.

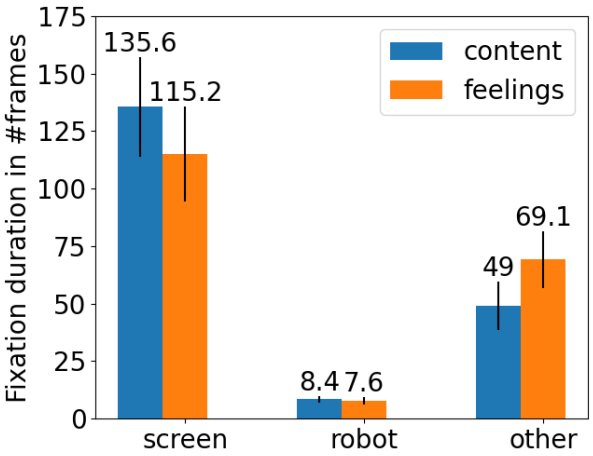


Figure 4.13: The average duration of the fixations on each visual target, aggregated per scenario and grouped by the labels memorized because of content and memorized because of feelings. The mean and 95% confidence interval are displayed in black.

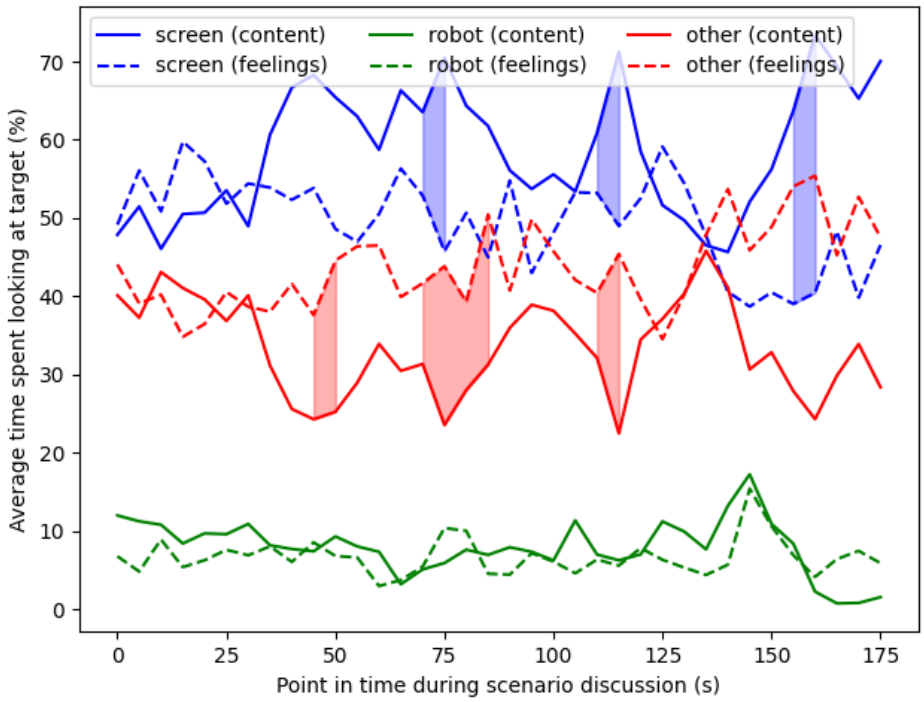


Figure 4.14: The average time spent looking at the 3 different visual targets throughout the duration of a scenario discussion, in steps of 5 seconds and grouped by the reason-labels content and feelings. The shaded areas are the intervals in which a Mann-Whitney U test found significant ($p < 0.05$) differences between the two groups (content vs feelings) for each visual target.

discussions ($t > 45$ seconds).

4.3.3. Machine learning model

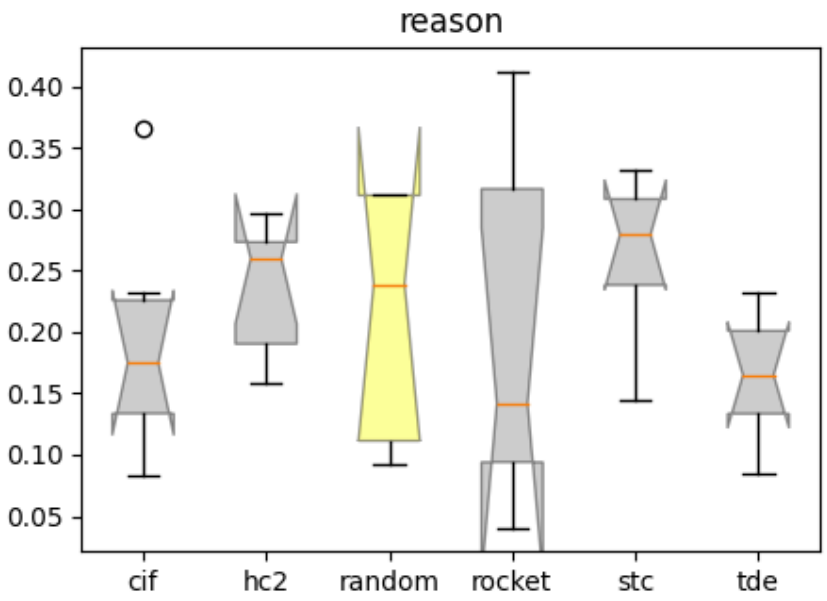


Figure 4.15: Boxplots of the performance of each model for the reason dataset. The models that performed significantly better than the dummy classifier (highlighted in yellow) are highlighted in blue.

In order to verify whether the patterns found in this dataset can be automatically detected, the different machine learning models, DrCIF, ROCKET, STC, TDE and HIVECOTE-v2 were trained on the binary dataset. The results of this are plotted in Figure 4.15. In order to meaningfully analyze the model's performance, they are compared with the performance of a random (dummy) classifier through independent t-tests.

There was not a significant difference in the F1 score for problem 'Reason dataset' between DrCIF ($M = .214, SD = .11$) and Dummy ($M = .194, SD = .1$); $t(12) = .34, p = .741$. There was not a significant difference in the F1 score for problem 'Reason dataset' between ROCKET ($M = .214, SD = .11$) and Dummy ($M = .198, SD = .155$); $t(12) = .209, p = .839$. There was not a significant difference in the F1 score for problem 'Reason dataset' between STC ($M = .214, SD = .11$) and Dummy ($M = .263, SD = .068$); $t(12) = .92, p = .379$. There was not a significant difference in the F1 score for problem 'Reason dataset' between TDE ($M = .214, SD = .11$) and Dummy ($M = .164, SD = .055$); $t(12) = 1.012, p = .335$. There was not a significant difference in the F1 score for problem 'Reason dataset' between HIVE-COTE v2 ($M = .214, SD = .11$) and Dummy ($M = .237, SD = .059$); $t(12) = .437, p = .671$. In summary, for problem 'Reason dataset', none of the models perform significantly better than the random model.

4.4. Quality dataset

The quality dataset consists of 196 moments during the cHRIs that are classified as being either not memorized ($n = 56$) or memorized with low ($n = 39$), medium ($n = 36$), or high ($n = 18$) level of detail according to Table 3.1. This means there are a total of 4 classes in this dataset (see Figure 3.10).

4.4.1. Memory quality

Hypothesis 5: Children who can reproduce 'high quality' recollections of a scenario spend more time looking at the screen during these scenarios and less at the other visual targets, compared to children who reproduce 'low' or 'medium' quality recollections.

The total time distribution over the visual targets can be seen in Figure 4.16. An ANOSIM test showed that there is no significant main effect between the two groups ($p = 0.997$). However, in the chart there appear to be significant differences in the time spent looking at each target. This suspicion is

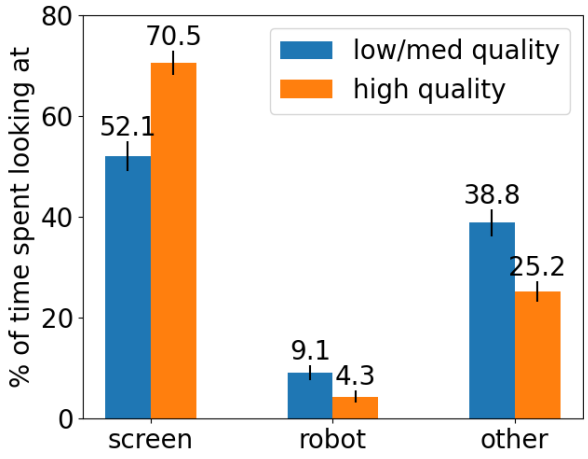


Figure 4.16: The relative time spent looking at the 3 different visual targets, aggregated per 30 seconds and grouped by the labels low/medium quality memorized and high quality memorized. The mean and 95% confidence interval are displayed in black.

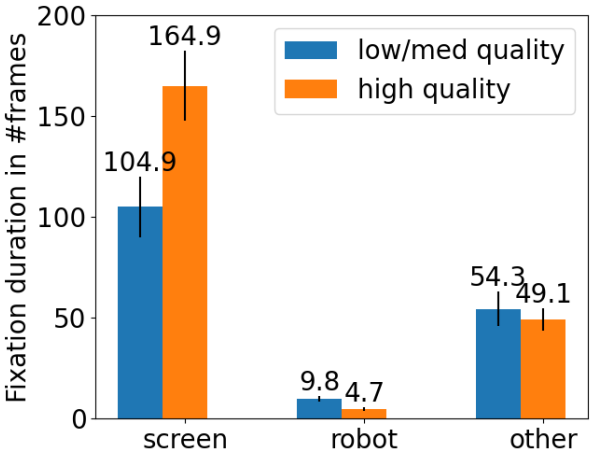


Figure 4.17: The average duration of the fixations on each visual target, aggregated per scenario and grouped by the labels low/medium quality memorized and high quality memorized. The mean and 95% confidence interval are displayed in black.

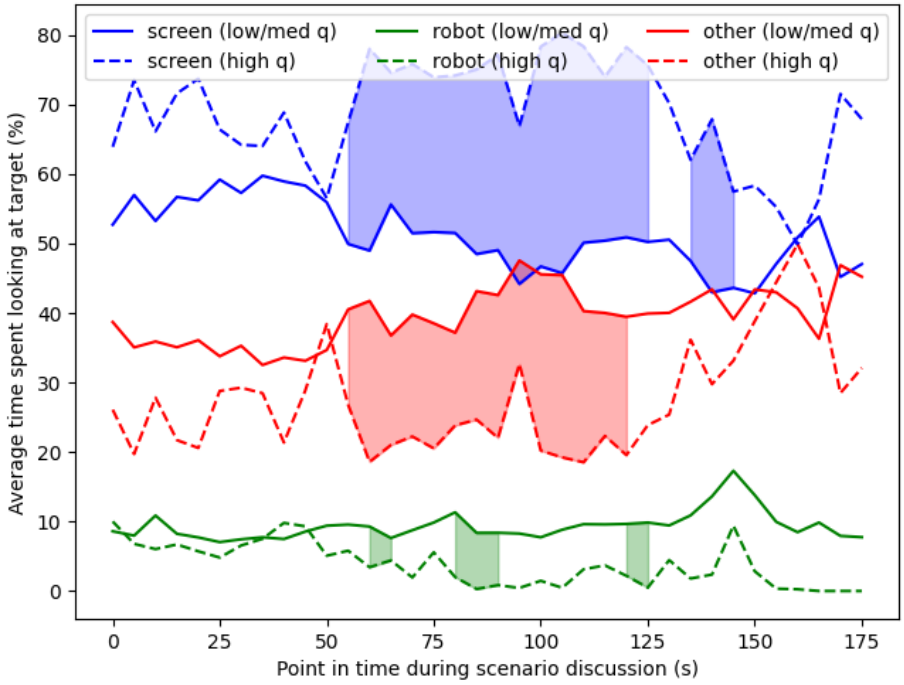


Figure 4.18: The average time spent looking at the 3 different visual targets throughout the duration of a scenario discussion, in steps of 5 seconds and grouped by the labels low/medium quality and high quality. The shaded areas are the intervals in which a Mann-Whitney U test found significant ($p < 0.05$) differences between the two groups (low/med q vs high q) for each visual target.

confirmed by Mann-Whitney U tests for screen ($MWU = 17179.5, n_1 = 454, n_2 = 105, p < 0.001$, Effect size by rank-biserial correlation: small); robot ($MWU = 30433.0, n_1 = 454, n_2 = 105, p < 0.001$, Effect size by rank-biserial correlation: small); and other ($MWU = 29467.0, n_1 = 454, n_2 = 105, p < 0.001$, Effect size by rank-biserial correlation: small). These tests show that the children who can reproduce high quality recollections spend more time looking at the screen and less at the robot and 'other'. In addition, an ANOSIM test showed that there is no significant main effect ($p = .307$) between the two groups for the fixation duration on each target (Figure 4.17). However, Mann-Whitney U tests showed that the average fixation duration was significantly lower on the robot ($MWU = 24170.5, n_1 = 393, n_2 = 94, p = 0.0$, Effect size by rank-biserial correlation: medium) and higher on the screen ($MWU = 12648.0, n_1 = 393, n_2 = 94, p = 0.0$, Effect size by rank-biserial correlation: medium), for children who reproduced high quality memories, compared to low or medium quality memories. No significant difference was found in the fixation duration on 'other' ($MWU = 17877.0, n_1 = 393, n_2 = 94, p = 0.628$).

The gaze behaviour throughout the scenario is plotted in Figure 4.18, where the shaded areas display the intervals in which the difference between the two groups, low/medium quality memories and high quality memories, is significant for each visual target. The differences in gaze behaviour between the different qualities of recollections appear to be most significant in the middle part of the scenario discussion ($55 < t \leq 140$).

4.4.2. Machine learning model

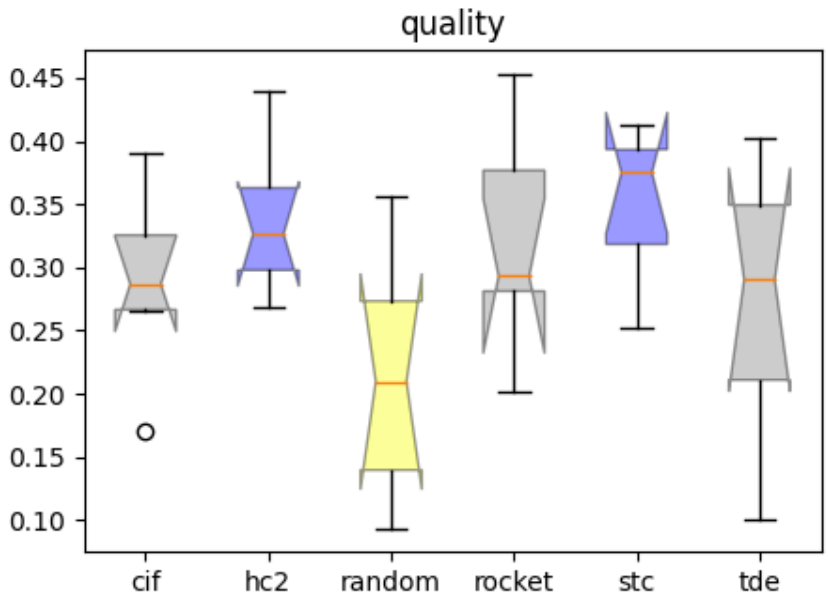


Figure 4.19: Boxplots of the performance of each model for the quality dataset. The models that performed significantly better than the dummy classifier (highlighted in yellow) are highlighted in blue.

In order to verify whether the patterns found in this dataset can be automatically detected, the different machine learning models, DrCIF, ROCKET, STC, TDE and HIVECOTE-v2 were trained on the binary dataset. The results of this are plotted in Figure 4.15. In order to meaningfully analyze the model's performance, they are compared with the performance of a random (dummy) classifier through independent t-tests.

There was not a significant difference in the F1 score for problem 'Quality dataset' between DrCIF ($M = .213, SD = .099$) and Dummy ($M = .289, SD = .074$); $t(12) = 1.493, p = .166$. There was not a significant difference in the F1 score for problem 'Quality dataset' between ROCKET ($M = .213, SD = .099$) and Dummy ($M = .321, SD = .092$); $t(12) = 1.947, p = .08$. There was a significant difference in the F1 score for problem 'Quality dataset' between STC ($M = .213, SD = .099$) and Dummy ($M = .352, SD = .062$); $t(12) = 2.904, p = .016$. There was not a significant difference in the F1 score for problem 'Quality dataset' between TDE ($M = .213, SD = .099$) and Dummy ($M = .273, SD = .112$); $t(12) = .976, p = .352$. There was a significant difference in the F1 score for problem 'Quality dataset' between HIVE-COTE v2 ($M = .213, SD = .099$) and Dummy ($M = .337, SD = .062$); $t(12) = 2.594, p =$

.027. In summary, for problem ‘Quality dataset’, only models STC and HIVE-COTE v2 perform significantly better than the random model.

5

Discussion

This chapter will discuss and review the results from chapter 4. In addition, the findings will be put in scientific and societal real-life context to discuss their implications and relevance in relation to children's health and education. More specifically, section 5.1 will reflect on the labels acquired through thematic analysis and compare them to existing literature, section 5.2 will reflect on the machine learning experiment, section 5.3 will reflect on the data analyses and relate them to relevant literature, section 5.4 will identify the limitations of this study and section 5.5 will put the findings into context.

5.1. Labelling

Following the method described in chapter 3, different labels, i.e. reasons for remembering moments, were identified from the children's post-interaction interviews. As stated before, there is limited literature available on memorability of moments and thus for reasons for remembering. Nevertheless, the work by Tsfasman et al., 2022 includes a similar study and also identified main themes, or reasons, for remembering, which are listed in Table B.1. In this section, these themes will be compared to the ones found in this study.

Similar to the work in Tsfasman et al., 2022, participants indicated reasons for remembering moments such as *time* ('it was the last scenario') and *shared_experience* ('I have experienced the same thing'). On the contrary, however, the participants did not indicate that any behaviour of the other party (i.e. the robot) was a factor for their memorization. At times, they would make general comments about the robot ('He didn't listen very well.'), but these were not tied to specific moments. This is probably due to the fact that the robot, unlike the humans in Tsfasman et al., 2022, was not moving or showing emotions.

The labels *fact_about_others* and *fact_about_world* from Tsfasman et al., 2022 are loosely related to the label *content* in this study. However, in this study, the topic of the conversations are more directed, limited and subjective, allowing for less unexpected information or facts to be introduced. In addition, the scenarios were set up so as to be easily understandable and familiar for the children, so they are not caught off guard by unknown information. Nevertheless, the content that was discussed was often still part of the reason for remembering. Participants indicated multiple times, for example, that they remembered the scenario due to it 'never happening before' or because of the appearance of a 'sandwich' or 'flute' in the stories and/or visuals on screen.

Moreover, the label *empathy* in this study can be related to both the *cognitive* and the *self_perception* label in Tsfasman et al., 2022. These are two separate labels in the latter because adults are generally able to distinguish clearly between their own feelings and relating to another. Children, on the contrary might have more difficulties with compartmentalizing all these feelings. Another reason why it is hard to distinguish between the participants' own feelings and feelings of empathy is that, by design of the interaction, the children have to look at different situations from a fabricated perspective and act as if they are in that situation. This means that their feelings of empathy for people in that scenario would actually be their own feelings, given that they would be placed in that scenario. See Table B.2 for an example. Whereas the participants in Tsfasman et al., 2022 are expressing empathy towards their conversational partners, the participants in this study empathize with fictional people in fictional

scenarios, which makes all feelings of the participants more inward and blended together.

Finally, this study required the creation of two new labels, *thought_&_ability* and *no_reason*, that did not occur in Tsfasman et al., 2022. The presence of the *thought_&_ability* label was needed because many participants indicated that they remembered a certain scenario because they found the content easier to process, e.g. in Table B.3. In addition, by design of the interaction, the children are required to make decisions and there are cases in which they indicate that the difficulty or ease of making that choice is their reason for remembering. Generally, adults do not have difficulties following a 'normal' discussion, so this logically would not have been a reason for remembering a moment in Tsfasman et al., 2022.

The label *no_reason* was necessary to be introduced because many children had difficulties identifying a reason for remembering a moment. In 42% of all self-identified memorable moments, the children were not able to formulate why they remembered this moment. Whereas adults can typically meta-analyze their own behaviour and come up with something, for children this type of thought-behaviour might be too conceptual. In addition, during the interviews it appeared as though some children were getting shy or insecure during this 'why' question. This could be due to the fact that it is uncommon to ask for a reason for remembering something and they were not sure how to provide a proper answer or answer that would be seen as 'correct' by the interviewer. The thematic analysis process made it possible to meta-analyze the language used by the participants and identify hidden reasons for remembering through 'coloured' or motivated memory descriptions. For example, in Table B.2 the participant remembers that they motivated their choice to increase the happiness of everyone involved in the scenario, which is related to empathy. These 'indirect' motivations were annotated throughout the dataset according to the same labels as the 'direct' reasons for remembering and they serve as a backup label in case the child says to have no reason for remembering that moment. Nevertheless, sometimes there is no strong enough evidence for such an indirect label, and the final label remains 'no reason'.

5.2. Classification

5.2.1. Results per problem

Overall, the highest weighted F1 scores for each problem are as follows: 0.263 for the reason dataset (1.23×0.214, the dummy classifier's score), 0.603 for the binary dataset (2.89×0.209, the dummy classifier's score), 0.468 for the handpicked dataset (1.45×0.322, the dummy classifier's score) and 0.352 (1.65×0.213, the dummy classifier's score) for the quality dataset. This means that, relatively, the order in which the models' scores improved most over the dummy classifier is: the binary dataset, the memory quality dataset, the handpicked binary dataset and the reason dataset.

Notably, the handpicked binary dataset produces lower test scores than the full binary dataset on a very similar (binary) problem setting. This is not necessarily as expected, since, by definition, the handpicked dataset should contain data of higher quality, i.e. there should be less noise/inaccurate gaze target flags and this should lead to more similarity within the classes and higher separability between the classes. A likely explanation for this, is the fact that the handpicked binary dataset is smaller than the full binary dataset (78 vs 149 samples). As such, the handpicked dataset is critically small, meaning that the expected gain in data quality can not make up for the loss in data quantity.

The fact that the binary labeled dataset results in the highest scores is in line with expectations. Namely, given that there are only 2 possible outcomes, memorable or not memorable, the problem allows for the most amount of training data per class. In fact, the order of performance in the different problems is according to the number of classes in the problems: 2 in the binary and handpicked binary datasets, 4 in the quality dataset and 5 in the reason dataset. Again, a likely explanation for this is the difference in the number of samples. If the same dataset is divided into more classes, it means that there are less samples per class and therefore less (reliable) evidence for each class definition.

Nevertheless, the difference in performance between the reason dataset (maximum F1 score of 0.263) and the quality dataset (maximum F1 score of 0.352) is relatively big, considering they have a difference of only 1 in the number of classes. This suggests that the separability between the quality classes is bigger than that of the classes in the reason dataset. A reason for this could be that eye gaze patterns are more indicative of the quality of the memorability of the moment than the motivation behind the memorability of the moment. Another reason could be that separability of the classes in the reason dataset is small because the children were not able to adequately self-identify their reasons for remembering. This was observed through the large number of "I don't know" answers to the

question of why they remembered a moment. This also lead to the fact that part of the reason dataset was annotated, not with the children's self-identified reasons, but by reasons deduced from thematic analysis. Finally, the difference in performance could still be due to the fact that there are more classes present in the reason dataset. Even more so because the training data is undersampled according to the minority class and, as such, the models might have had too little training data to identify reliable patterns in the reason dataset.

5.2.2. Results per model

For each problem setting, the model that obtains the highest combined balanced accuracy score and F1 score, is STC. It can therefore be concluded that STC, or, more generally, a shapelet-based time series classification algorithm, is the model architecture most well-suited to data of this type, i.e. long sequences of multi-dimensional categorical data. Whereas all the other models in this study use summary statistics and more generally extracted patterns to classify samples, STC directly uses the distance between raw subsequences of samples (Bagnall, Flynn, Large, Lines, et al., 2020; Lines et al., 2012). This indicates that, for this dataset, class separability can best be found on a micro-scale, i.e. the classes are defined by sub-patterns in the time series. In addition, other models use counts of a relatively small selection of similar patterns to decide class-correspondence, whereas STC uses a larger number of patterns and rather their presence in samples than the count (repeated presence) (Bagnall et al., 2018). Again, this shows that classes are more easily separated according to numerous smaller differences, as opposed to more general similarities. Finally, it can be argued that the distance vectors generated by STC maintain more of the sequential characteristics of the original dataset than other models that summarize or compress the data, to process it more like a batch. This goes to show that appreciating the authentic structure of data is valuable when extracting meaning from it.

On the downside, STC is computationally expensive compared to the rest of the models. STC and HIVE-COTE v2 are slower than DrCIF, ROCKET and TDE. Both ROCKET and DrCIF are exceptionally fast and do not fall far behind in terms of test scores, which is in line with findings in literature reviews (Bagnall et al., 2016; Bagnall et al., 2018; Dau et al., 2019). Moreover, especially ROCKET appears to perform very well in settings with fewer data (e.g. the handpicked binary dataset). Because of these two characteristics, ROCKET makes for an excellent benchmark algorithm (for testing). In more data-rich settings, however, ROCKET can not match up to a more specific and elaborate algorithm like STC. DrCIF on the other hand, appears to be performing better on larger datasets than on smaller ones. In general, it would be expected that all models would benefit from a dataset with more samples, but especially DrCIF and TDE (and therefore HIVE-COTE v2), since these algorithms might be able to find more meaningful summary statistics or batches in a larger pool of options.

HIVE-COTE v2 has been shown to perform very well in literature (Bagnall et al., 2016; Bagnall et al., 2018; Middlehurst, Large, Flynn, et al., 2021). In this study, however, its performance is somewhat on par with those of the other models. This could be due to the fact that HIVE-COTE v2 requires a lot of resources to train its members and is by default restrained by a maximum runtime, which prohibits the members to reach their full potential. This could, in fact, limit especially the performance of STC as a member of HIVE-COTE v2, since it requires the most time and space. So, even though HIVE-COTE v2 can be a good starting point, given accurate and applicable domain knowledge it might be better to select a model suitable to the domain and to fully exploit that model's potential, rather than limiting the performance.

5.3. Data analyses

This section will review the results with respect to the patterns found in chapter 4 and relate them to relevant literature.

5.3.1. Summary of results

Through statistical data analyses, it was found that there are significant differences between the gaze-time distribution over the visual targets ('robot', 'screen' and 'other') in a number of settings. First of all, between the not memorized moments and the moments memorized with high or medium quality. In the memorized case, participants look significantly more at 'robot' and 'other' and less at 'screen'. Secondly, there is a significant difference between moments that are memorable due to the cognitive ease and moments that are memorable because of other reasons. Specifically, participants that remember

moments due to ease look significantly more at ‘robot’ and ‘other’ and less at ‘screen’, compared to moments remembered because of other reasons. Lastly, there is a significant difference between moments that are memorized due to the content and moments that are memorized due to related feelings. Post-hoc tests revealed that participants who remember moments due to feelings rather than content spend less time gazing at ‘screen’ and more time gazing at ‘other’.

Tests that compared the gaze alternation patterns found that the number of within-screen gaze alternations is significantly lower for the memorized case, compared to the not memorized case. This difference is especially apparent in the first 60 seconds of the scenario discussions.

Additional tests found no significant main effect, but deeper analysis found some noteworthy patterns nonetheless. Particularly, participants that said to remember a moment due to the cognitive ease, as opposed to other reasons, were shown to have shorter fixations on ‘screen’. In addition, participants that were able to reproduce high quality memories, as opposed to low and medium quality, were shown to have longer fixations on ‘screen’ and short fixations on ‘robot’.

Finally, research into the gaze-time distribution over time throughout the scenario discussions also identified common patterns. Namely, the comparison between the memorized samples and not memorized samples showed that the significant difference between the visual targets is the most apparent in the first $t \leq 30$ seconds, which is roughly equal to the time frame in which the robot introduces the scenario by means of a monologue ($t \leq 35$ seconds). Moreover, the differences between types of memorized moments were found to be most significant in the middle and ending of the scenario discussions. In particular, the temporal analysis showed that the significant differences between reason labels ‘feelings’ and ‘content’ occur from the middle towards the end of the scenarios ($t > 45$ seconds). It was also shown that the significant differences between the visual target distributions of the low and medium quality classes compared to the high quality class were found in interval $55 < t \leq 140$ seconds.

5.3.2. Implications of the findings

Robot-related gaze It has been established that, compared to not memorable moments, children tend to look more at the robot, i.e. the conversational partner, and at miscellaneous points in space during memorable moments. The fact that the children look more at the robot during memorable moments is somewhat contradictory to the findings by Tsfasman et al., 2022, which state that people look less at other people during high levels of memorability. This can possibly be explained by the fact that the conversational partner in this experiment is a (single) robot, instead of multiple other humans. This might suggest that the children are not engaging with the robot in a similar way as they would with humans. According to literature, gaze aversion from the conversational partner can signal the encoding of information, intimacy modulating behaviour or turn-taking behaviour (Admoni & Scassellati, 2017; Doherty-Sneddon & Phelps, 2005; Glenberg et al., 1998; Oertel et al., 2012). Since the robot is not showing any bodily or facial expressions, turn-taking signals and intimacy modulating signals are nonexistent from the robot’s side. It has also been shown that gaze aversions that signal cognitive effort are longer (Andrist et al., 2013). So, for this study, it is more likely that children look less at the robot during memorable moments to process these moments and encode the (large amount of) incoming information. An additional explanation for the discrepancy between the expectation and the observed behaviour is likely to be found in the inherent differences between human conversational partners and robot conversational partners. Namely, the purpose of this gaze aversion is to limit the information intake and make cognitive space for mental processing. However, the NAO robot’s appearance does not provide any additional information regarding the content. Specifically, as opposed to humans, the NAO robot in this experiment does not show mouth/lip movements while talking, does not move its limbs or body and keeps its head/eyes fixated on the child nonstop throughout the interaction. This means that, as there is no information overload from looking at the robot, the children would be less inclined to avert gaze from the robot during cognitively demanding moments, compared to humans. Moreover, given the lack of gaze-related, or other body language-related, feedback from the robot that evoke typical social response behaviour, the child might be more inclined to gaze or stare at the robot to, for example, try to gather these cues (in vain).

The work by Tsfasman et al., 2022 is the only known research on the relationship between gaze and memorable moments. Nevertheless, there is a broader range of literature on gaze patterns for related measures like engagement and attention, that also specialize in gaze patterns during HRI. According to Rich et al., 2010, the intention behind mutual facial gaze is to maintain engagement and Nakano and Ishii, 2010 states that looking at a conversational partner is a signal of engagement. It

could be stated that humans who are more engaged are absorbing more (topic-relevant) information. Intuitively, the intake of information correlates with information reproduction (i.e. memory recall). This would explain the heightened gaze-time at the robot for moments that were remembered, compared to moments that were not. As stated before, this effect might be present/bigger in this study compared to others, because the robot features less human characteristics that could lead to children's gaze aversion because of implicit social rules. Nevertheless, perhaps this points to be a possible positive correlation between children's engagement in a moment and the memorability of this moment.

The results have also shown that children tend to look away from the robot during moments that they identified as memorable due to their personal feelings. On the contrary, it was found that children look more at the robot during moments they experienced as cognitively easy. This could be an indicator of the fact that moments of deep thinking on morals or feelings are perceived as cognitively difficult, which would explain gaze aversion (Breil & Böckler, 2021). However, another explanation could be that the children are not comfortable to share the more emotional moments with the robot, possibly due to a dislike for the robot or a lack of trust. In literature, gaze aversion is indeed related to negative social cues (Burgoon et al., 1986). While the evidence on the relationship between cognitive load, gaze aversion and memorability are consistent throughout the results in this study, further research is needed to rule out the possibility of a superficial social bond between child and robot.

Whereas, generally during memorable moments, children look more at the robot, the results show that children fixate for significantly shorter periods of time on the robot during high quality memorable moments compared to low or medium quality memorable moments. A possible explanation is that, during high quality memorable moments, children avert their gaze to process the information, but then regularly check in with the robot to indicate engagement. This theory is supported by Admoni and Scassellati, 2017, who state that people tend to look away from a partner shortly and frequently to modulate intimacy and engagement. This means that the children who remember a moment well regularly signal their understanding and engagement with the conversation to the robot, as they would with a human conversational partner. Longer fixations have also been correlated with mind-wandering (Bixler & D'Mello, 2016). Perhaps children are also able to reproduce higher quality memories when they spend less time mind-wandering, indicated by shorter fixations. However, this is a less likely explanation, since people tend to fixate on more spread out or still visual targets during mind-wandering, rather than on faces (Hutt et al., 2019).

In summary, there appears to be an overlap or positive correlation between gaze-based signs of engagement and memorability. However, engagement between a human and a robot might be expressed differently from engagement among humans. As expected based on literature reviews, children are more likely to remember moments if they show signs of relatively deep thinking, like gaze aversion, during these moments.

Screen-related gaze According to the results, children look less at the screen, which shows visual aids to the topic of conversation, and more at 'other' during memorable moments in cHRIs compared to not memorable moments. Gaze aversion, away from a source of information, has been linked to cognitive processing in literature (Doherty-Sneddon & Phelps, 2005; Glenberg et al., 1998), especially for children (Doherty-Sneddon et al., 2002). As such, it seems again that children are more likely to remember moments of deep thinking, which would be in line with the findings in Tsfasman et al., 2022.

It seems as though the information on screen can be too much to process for the children. This is especially apparent in the temporal analyses, in which the children who remember the moment are averting eye gaze from the screen especially in the beginning, which is when they are verbally provided with all the information regarding the scenario to be discussed. This finding is also in line with early research on gaze aversion due to cognitive overload (Cegala et al., 1979). It also explains the significant increase in looking at gaze target 'other' during memorable moments, as gaze during gaze aversion is logically being turned towards 'other'. Linked to this, are the results that showed that children who remember a moment because of feelings or moral considerations show this behaviour (looking less at 'screen' and more at 'other') significantly more than children who remember moments for other reasons. Arguably, these children are most invested into the moral dilemma they have to solve and put the most cognitive effort into finding a resolution. This cognitive effort is also shown once more to be positively correlated with memorability, as feelings or morals are provided as the most prevalent (34%) reasons for remembering, as indicated by the children themselves.

Children were also shown to have longer fixations on screen and less gaze alternations within the

screen when they were able to reproduce a moment with high quality, as opposed to low or medium quality. Humans have been shown to fixate longer on images during explicit tasks like comparison (Loftus, 1972), and these longer fixations are also indicative of more efficient cognitive effort in these tasks (Eckstein et al., 2017). In addition, in research on binary value-based decision-making, it was shown that fixation duration tends to be higher for problems considered 'difficult' (Krajbich et al., 2010). This relation is also supported by the results of this study that showed that children who remember something due to cognitive ease have shorter fixations on screen, compared to children who remember moments for other reasons. Again, the general trend shows a positive correlation between the cognitive difficulty experienced during a moment and that moment's memorability. However, as mentioned above, a longer fixation duration has also been linked to mind-wandering. It is important to keep in mind that these longer fixations are present in the subset of data that has lower screen-gaze time overall. This indicates that these fixations in the memorable case are not present due to mind-wandering, but are instead carefully and intentionally directed for information intake.

Finally, children who indicated to remember moments due to a notable aspect of the (visual) content were shown to gaze more at the screen (and less at 'other') during these moments. So, within the group of memorized samples, in which gaze-time at the screen is less than in not memorized samples, gaze-time at the screen is increased for children who show particular interest in an item, word or character that is likely displayed. This trend is in line with literature that states gaze fixation on an object is indicative of interest in that object Hirayama et al., 2010. As such, gaze detection can differentiate between children that are interested in the moral or emotional aspect of a dilemma and children that are interested in specific content.

Temporal effects The temporal analyses of the gaze distribution over the visual targets showed that the distinction between memorable moments and not memorable moments is made mostly in the the first 30 seconds of the scenario discussion. This corresponds directly with the time period in the discussion during which the robot uninterruptedly explains the moral dilemma and the available options to choose from, and the screen changes its content to the visuals relevant to this scenario. This information serves as the baseline for the memory that the children are asked to recall during the post-interaction interview. Following the results, it seems vital that the children are encoding these first 30 seconds into memory. This encoding process is characterized by gazing less at the screen and more at 'other', for reasons explained above. Beyond these 30 seconds, there appear to be no more moments within a scenario discussion that are significant indicators of memorability. This shows that the gaze behaviour at a specific point (or 'window') in time is directly indicative of whether that specific content will be remembered or not.

For this experiment, all the factual information that serves as the foundation for the scenario and thus the memory is presented at once in the beginning. It is necessary for the child to store all of this information, at least in short term working memory, in order to make a value-based decision between the two options. From the temporal analysis it seems that if the child manages to do this encoding successfully, and thus stores the information in memory, they can continue to think deeper about the problem. The temporal analyses have shown that the distinction between different reasons for remembering and the quality of the memory is most easily distinguished in the middle and end of the scenario discussion (> 45 seconds). It appears as though the child's internal monologue, i.e. a value-based dilemma as in the presented scenario, serves as a selection and filter process of what should be prioritized in memory. The gaze pattern analysis has shown that it is possible to determine whether children who remember a moment are especially mindful of, for example, the thoughts or feelings behind the moral dilemma, or rather notable facts about the setting in which the dilemma takes place.

In addition, the amount of details that a child remembers from the scenario is also determined in the second part of the scenario discussion. The analysis showed that the first 30 seconds should provide enough information to the children to allow them to at least recognize the same problem from memory, but the children require a more longitudinal study process to reproduce the moment with a lot of details and to add their own meaning to them or internalize them.

5.4. Limitations

This section will identify limitations of the study performed. These consist of pre-set constraints, as well as limits identified during the process of the study. All aspects of the study will be discussed.

5.4.1. cHRI

The cHRI was originally designed to last $20 - 30$ minutes, in which 8 different scenarios were to be discussed. In practice, this proved impossible and after the first 4 participants it was decided to cut the number of scenarios to be discussed down to 4, and thus to hold 2 separate sessions of 4 scenarios instead. The reason for this was the fact that the ASR algorithm in the NAO robot was insufficiently adjusted to understanding children. There is limited research available related to ASR for children, and even fewer public (open source) applications, especially in dutch. Additionally, using ASR algorithms by big, corporate players in the AI and ASR field is not possible as they are cloud-based and therefore not in correspondence with the data security protocol. The effects of this on the cHRI do not only have to do with the speed of the interaction, but also with the quality. For the participants, the inconsistencies and failures of the robot lead to annoyance and reduced excitement or willingness to cooperate in the following sessions. Initially, most children were excited to interact with the robot, but the robot was subsequently not able to meet their expectations. For example, the robot often misunderstood their answers, did not respond at all, or mispronounced their names. As for the research, the ASR issues were reason for a lot of interventions by the cHRI moderators, who would repeat instructions to the child regarding the way of speaking, or who would repeat the children's answers in a more ASR-proof way. This not only disturbs the natural flow of the cHRI, but also causes the children to turn towards the moderator instead of the robot and thus show diverging eye gaze patterns, that were not accounted for by the eye gaze algorithm or visual target identification.

Nevertheless, the adjustment from interactions with 8 scenarios as compared to 4 scenarios might have been beneficial for the results regarding memory. During the post-interaction interviews it was found that children who took part in interactions with 8 scenarios had a lot of difficulties to reproduce any of the scenarios, whereas the ones who treated only 4 scenario usually could reproduce at least one. When researching children's memory, it appears to be important to not overload them with information, as it could cause them to shut down.

The cHRIs were held consecutively for 3 days with limited breaks. The NAO robot has problems with staying 'active' for such long periods on end. For this reason, it was decided to not include any movements by NAO in the interaction. In fact, the NAO would be standing during its introduction speech to the child, but during the scenario discussions it would sit down. In addition, by design, the NAO robot lacks any facial expressions and is 'stuck' in a sort-of curious, open-minded gaze, directed at the person who it is interacting with. Usually, this would be the child, but at times, due to the ASR issues, it would even direct its head/gaze at the moderator instead of the child. As a result of all of this, the NAO robot was not showing any signs of non-verbal communication or feedback towards the child and its functions could have been executed by a loudspeaker as well. The children are likely not aware of this and attribute human-like qualities to the robot regardless. Nevertheless, this could have influenced the way that the children bond with the robot or cause them to relate to the robot negatively; something that could be improved in future studies.

Finally, the cHRI performed for this study, also served a purpose for a multitude of other studies. Therefore, at times there were unaligned interests and priorities of the people involved in setting up the experiment. In addition, for the moderator, there were a lot of tasks and steps to take before starting and finishing each interaction. As a result, the children were not always carefully and systematically instructed to try and sit (relatively) still, to allow the gaze tracking algorithm to function more easily. This is also due to the fact that it was slightly underestimated how much some children move around while sitting on a chair.

5.4.2. Post-interaction interview

During the post-interaction interview, the goal was to find the ground truth identification of memorable moments and not memorable moments. These findings were reported in chapter 3 and chapter 4, but the post-interaction interview also provided some qualitative insights into the children's behaviour and thought process. First of all, notably, some children were able to remember/reproduce all scenarios discussed and some children none. The children that remembered none were often younger and less communicative in general. It is therefore not always believed that children who say they do not remember anything actually do not remember or if they simply refrain from saying anything. Perhaps these children would benefit from a different interview style or method to better express their thoughts. Given more resources, some children could be given more time to remember, or could be given pen and paper to collect their thoughts privately.

If needed, the children were provided help to remember: a reproduction of the visuals that appeared on screen during the different scenarios. This proved helpful for a majority of the children, who otherwise wouldn't have remembered any of the scenarios. During classification and data analysis, however, limited distinction was made between spontaneous memories and those that required triggers through visual cues. The reasoning behind this was that the topic of interest in this study is not to detect the type of memory use or recall, but rather a memorable moment no matter how it is remembered. It is unclear if the children actually needed the visual aid to recall memories, or if it gave them a better understanding of the recall task and/or if it took away insecurities regarding speaking up about their memories. This relates back to the point made above regarding shyness. In a future study, perhaps the introduction of visual aid could be more structurally organised to enable a comparison of the different recall methods

5.4.3. Eye gaze tracking

Even though the design of the algorithm for eye gaze tracking was not included in this study, its operation does affect the results of this study. The eye gaze tracking algorithm was specifically adjusted for this experiment and has been set up such that it is optimized for the camera angle that it was in and also to work well with children. However, only adjusting the algorithm to work with children's head and face shapes is not taking into account the differences in behaviour between children and adults. Namely, in practice, the algorithm was struggling to adapt to the children's peculiarities in movements. Some children were quite restless and changing their seating or head position almost constantly, and the visualization of the gaze tracking algorithm showed that the gaze calculation for these children/moments is often incorrect. In addition, the gaze tracking algorithm is frequently incorrect when dealing with participants with glasses. This indicates that the field of eye gaze tracking is still in development and still needs to be improved in flexibility, especially with regards to child subjects.

5.4.4. Memorable moments classification

A discussion of the results of the machine learning models for gaze-based memorable moment detection in this study is included in section 5.2. Nevertheless, it should be noted that the model comparison did not include deep learning models. As deep learning approaches have been making their way into the time series domain, they have challenged current state of the art models and are rapidly improving and increasing in popularity (Fawaz et al., 2019). In the future this study could be repeated with the addition of one or more deep learning time series classifiers, such as state-of-the-art ResNet (Z. Wang et al., 2016), to compare their performance and perhaps achieve better results. A precondition for this to be feasible, however, is that there should be a lot more data available to train the models. Deep learning architectures have more parameters to train and optimize than standard machine learning models and thus require more training data. Perhaps the data collected in this study could be the start of a new knowledge base regarding eye gaze data during cHRI and future studies can add to this.

In order for one of the presented models to be applied in practice, a desirable feature would be the inclusion of a confidence level over the predictions. If provided by the gaze tracking algorithm, the confidence over the eye gaze precision could also be included. Such a feature would allow for targeted testing with individuals that score a high confidence, as well as targeted points of improvements for individuals with low confidence scores.

5.4.5. Identification of heuristics

This study does not only present the performance of different classifiers on memorable moments, but also attempts at finding heuristic patterns for child eye gaze behaviour during memorable moments. This analysis is valuable for three reasons. The first is that it validates the data and the class divisions: finding systematic differences between classes shows that they are likely to have some common meaning. The second is that it may provide some insight into macros that the machine learning models are also extracted from the samples. The third is that it may inspire other researchers to develop case-specific (machine learning) algorithms for detecting memorable moments during cHRI with gaze, or new ways of extracting meaning from this type of data. Searching manually for patterns and common themes in large datasets, however, is a potentially endless process with no real substantial connection to patterns detected by machine learning models. In addition, these type of analyses are subject to the researcher and their area of focus. Reviews by other researchers or reproductions of this study are therefore very welcome and should further confirm the validity of the findings.

One of the important findings of this study is the identification and differentiation of longitudinal information processing in relation to memory. Two distinctive parts of the scenario discussions were identified ($t \leq 30$ and $t > 30$), in which the content discussion and mental processing appear to be different and to serve different processing functions. However, aside from the difference in content, there is also a difference in conversational structure between these two parts. In fact, during $t \leq 30$, i.e. the presentation of the dilemma, the child acts as a passive agent in the conversation, as the child does not provide any input during this time. The gaze patterns that were found during this moment are therefore not directly generalisable to gaze patterns that would occur during a more natural, conversational setting. In fact, they appear to be specific to a situation in which a child is presented with a lot of factual information (based around a moral dilemma).

5.5. Findings in context

Given the results from chapter 4 and their meaning and limitations as identified in section 5.2, section 5.3 and section 5.4, there is now ground to establish the depth of the implications of the findings in real life.

5.5.1. Scientific implications

Whereas there is a lot of existing research on humans' inner states such as engagement and attention, little is known about what makes humans define moments as memorable, as they are happening in the moment. Any research into the field of memorable moments detection would be a contribution on its own, but this study even more so because it is aimed at children, and during cHRI specifically, which are yet unexplored fields.

This study has shown that, with a relatively lightweight architecture, it is possible to achieve better-than-chance performing models for memorable moments detection. The model predictions are not infallible, but it is expected that the performance could be raised given a less noisy dataset and a hand-crafted model for the problem setting. It was found that a model based on local features, a shapelet-based time-series model, is the most well-suited for memorable moments detection. This is likely to be due to the fact it retains more of the sequential data characteristics than the other models and analyzes the data on a very low level, instead of with summary statistics. This gives reason to believe that gaze patterns during memorable moments are distinguishable from those during not memorable moments through the occurrence of very small sequential sub-patterns. Nevertheless, there were also differences to be found in patterns on a larger scale.

Generally, it was found that heuristics regarding eye movements related to engagement, attention and memorability can be used to reason over eye movement patterns of memorable moments. More specifically, children tend to gaze more at the robot during memorable moments, using short, frequent fixations. It is suggested that this is due to higher engagement and more regular modulation of mutual engagement/understanding. They gaze less, but in longer fixation sequences, at the screen providing information while listening to verbal information. This is thought to be due to cognitive processing, which has been linked to gaze aversion. In the moments after information provision, the children's (moral) reasoning and interests dictates their eye movements and influences their memory of the scenario. Children gaze more at other (non-labelled) objects in space during memorable moments compared to non memorable moments, again to facilitate information encoding and deeper thinking. Overall, children's thoughts on the topic of discussion influences what and how well they remember the discussion. Generally, the results have shown that deeper thinking leads to better memorisation, which is in line with related literature findings. Above all, the research in this study is of an exploratory nature and provides a lot of opportunity and reason for further and more in-depth research, which will be explored in chapter 6.

5.5.2. Societal implications

As this study was designed for cHRI, the findings can be used to improve the way robots interact with children. Whether it is through the incorporation of a machine learning model or with a rule-based implementation of the found heuristics, the findings from this study can be used to do live memorable moment detection during cHRI. The information regarding whether a moment is memorable or not can then be used in a multitude of beneficial ways. Firstly, on a short term, such models can be used to revise previously presented information. On the one hand, children could be excited with information that is already known to them and progress in a subject that is interesting to them. On the other hand,

learning processes can be improved by repeating the relevant non-memorable moments. Secondly, the robot could accumulate statistics on memorable moments and these could be used to highlight points of improvement or points of excellence in the cHRI, in a feedback loop.

Last but not least, creating a memory of memorable moments per each individual user could help to build up a profile of each child. The exact operation of this is left for future research, but it could be expected that long term modelling of children's memories provides insights into the interests and preferences of a child. This information can, in turn, be used to personalize the robot behaviour and the content discussed to the likes or needs, educationally or health-wise, of each child.

6

Recommendations

As this study is the first of its kind, and combines many recently developed technologies and ideas, there are a lot of lessons to be learned and takeaways from the experience. Chapter 5 has already highlighted some of the limitations and difficulties that were encountered in this study. This chapter will discuss how these obstacles can be prevented in future studies or reproductions and how the study and field of study can be augmented with additional, related research. More specifically, section 6.1 will highlight improvements to be made to reproductions or similar studies of this study and section 6.2 will discuss how research in related fields could also contribute to the progression of memorable moments detection using eye gaze during cHRI.

6.1. Reproductions or similar studies

Since this study is the first research into the relationship between children's eye gaze patterns and memorability, reproductions and similar studies would have a lot of added value in terms of validating and solidifying the findings. In order to enable smooth operations and usable results for these studies, this section highlights some guidelines and lessons learned regarding the setup of such a study.

6.1.1. Design

One limitation of the study at hand is the fact that 'memorable moments' are defined as 'memorable' after only testing for recollection immediately after the cHRI took place. Perhaps, after processing the events through continuing regular life or sleeping, the children would remember more, less, or different aspects of the interaction. A future study could take the post-interaction interview and hold it at a different time, e.g. one day later. Ideally, a study could do a long term effect analysis and hold multiple post-interaction interviews to model the memorability over time, e.g. immediately after, one day after and one week after. From an educational perspective, it would also be interesting to compare the memorability of teaching moments in cHRI to those in a traditional classroom setting (i.e. with a human teacher and surrounded by peers). Such studies could be useful, especially, to validate the finding that the eye gaze patterns of children are to some extent indicative of the quality or type of memory, as found in chapter 4.

6.1.2. Participants

Following the notes and findings on the children's behaviour in both the cHRI and the post-interaction interview, as highlighted in section 5.2, it is apparent that some children are more suitable for obtaining usable results than others. It should be clear that a study like this should be all-inclusive and representative of the population, but it should also be achievable to gather data, otherwise there are no results to be found or conclusions to be drawn. Especially at this early stage of research development in this field, obtaining baseline usable information is beneficial and a step in the direction of a better representation of society. Future studies could try to find participants that increase the chance of usable results by taking into account a participant profile, that was formulated after completing this study. The profile can be found in Table 6.1.

Participant attribute	Reasoning
is ≥ 9 years old	The younger participants were less capable of formulating their thoughts and memories in the post-interaction interview. They also showed more of some of the less desirable traits listed below, such as fear of the robot and shyness.
does not wear glasses	The eye gaze tracking software has been shown to have difficulties with recognizing eye gaze in participants with glasses.
has affinity with technology/robots and is, at least, not scared of robots	Certain children expressed discomfort when interacting with the robot and this should be avoided, as it is undesirable for the participants to feel this way and it might cause a negative bias in the results as well. It might also be interesting to investigate in a separate study how robots can be made to be less scary and more likable for children specifically.
does not have an attention deficit (hyperactivity) disorder	Whereas it might be interesting to study the effectiveness of cHRI for participants with a wide range of mental abilities and attention spans, this can better be studied in a separate, dedicated study, after establishing a baseline with proven reliability.
is open in communication and not overly shy	Some children were very quiet during the post-interaction interview and said to not remember anything. It is believed that this is not an accurate representation of their memory and is more likely due to shyness or unwillingness to cooperate.
is excited to participate in the study	Children who want to 'just get it over with' appeared to be answering out of convenience and according to interviewers' expectations rather than truth. This can lead to incorrect labelling.

Table 6.1: A non-binding profile with desirable qualities for participants in studies related to eye gaze and memorability in cHRI.

Some of the machine learning models in this study, especially in problem settings with several class labels, were likely limited in performance due to a lack of data (in the minority classes). If possible, this should be avoided in any future studies by maximizing the number of participants in the experiment. This is also a necessity if someone would like to add the inclusion of deep learning models to the comparison, since more data would be needed for training.

6.1.3. cHRI and post-interaction interview

Researchers looking to reproduce this study would be advised to make sure the cHRIs are of limited duration and information density. It was found in this study that a strict upper bound of 20 minutes and 4 scenario discussions (instead of 8) was beneficial to both the operation smoothness and quality of the results.

In order to enable a cHRI that is authentic and uninterupted by a moderator, the NAO robot should be equipped with better ASR technology. In any case, it is also advisable to create a fallback wizard-of-oz setting in case some part of the robot malfunctions or encounters problems in the practical setting. This would also ensure better quality eye gaze data, since the child would not be interacting with another conversational partner. chapter 5 also highlighted the limitations of the compatibility of the gaze tracking algorithm with children's behaviour in general. While better gaze tracking technology is being developed, future studies of this type can be more clear and persistent in instructing the children to sit still during the interaction.

During the post-interaction interviews it is advisable to make separate parts so as to analyze the data in a more categorical way. This could help to make a more clearly defined distinction between cued recall and free recall tasks as well, which would be an enhancement. This can be achieved by simply introducing a marker in the recording system, or using a fixed time per part. It could also be decided to leave out either the cued recall or free recall questions and focus on one of them, but this would be discouraged as it takes away some of the interview flexibility that is necessary to deal with the variety of children's personalities and behaviours.

Finally, it is desirable to verify whether the memorable moments detection is valid in different cHRI settings, not only when the children are making moral choices in fictional scenarios. It would be both

interesting and valuable to reproduce the study with a slightly different cHRI design and validate the results in this setting. An example could be a situation in which the robot is learning about children's dietary preferences and they have to choose between two different options a number of times.

6.1.4. Model

From the model comparison in chapter 4 it became apparent that some model architectures are more suitable for memorable moments detection in time series than others. This information provides some insights into the data characteristics and this gives reason to more speculation about what type of models or data extraction methods could lead to a better performance. As it became clear that STC was the highest performing model, it was concluded that the data is most accurately represented and thus classified through leveraging the occurrence of low-level sub-sequences within the individual samples. This information is useful and makes STC a good choice of model on its own. However, some potential information in the data is still unused. In particular, this is because STC, as well as the other models, is not leveraging the knowledge that the dataset comprises of eye gaze data, which, like any particular data domain, comes with its own peculiarities and characteristics. For example, eye gaze data is often analyzed based on fixations and saccades, and the data analysis in chapter 4 shows that there are non-accidental differences in these metrics between the different classes that could be used to find further separability. A custom application for memorable moments detection could be implemented to extract this information from the data and use it to train a model (or ensemble). The results of such a model could be combined with those of a local-features based model such as STC according through an appropriate voting mechanism. In summary, future research should dive deeper into appropriate machine learning architectures for this problem setting, given the findings in this study regarding suitable model types and data pattern heuristics.

It was also found during manual data analyses that the scenario discussions can be separated into 2 distinct parts in terms of patterns and functions. This knowledge can be leveraged to make dedicated models for each part and as such reach better and faster decisions. Given the large effect size of the difference in behavioral patterns for the memorable and not memorable moments in the first 30 seconds, a fast lightweight model could be trained on the binary case. A more advanced model could be used to distinguish the nuances in the memorized case during the second part of the scenario, without including the 'noise' from the not memorized case.

6.1.5. Multi-modal communication

Eye gaze is commonly used to study and reason over humans' inner state, but it is not the only available modality in conversational settings that can be leveraged for analysis. In Tsfasman et al., 2022, for example, memorable moments detection is performed using both eye gaze and audio/speech data. It was found that the speech activity in different moments has a correlation with the occurrence of a memorable moment. Therefore, future studies could also leverage the audio of the cHRI on top of the eye gaze data to possibly increase performance accuracy.

In addition, it is not unlikely that children's more general facial expressions and body language are also indicative of their inner state and whether they remember a moment or not. It will be left up to future research to identify all the different communication modalities that could be leveraged and how to combine them with the findings based on eye gaze in this study.

6.2. Recommendations for related research

Apart from studies that would reproduce or slightly alter this study, there are also related topics or fields of study or continuations that could aid in the progression of memorable moment detection during cHRI. One example that stems from the limitations found in chapter 5 is to research how robots can be made to be more child-friendly. This is deemed relevant because during a small portion of the post-interaction interviews, children admitted that they were uncomfortable with the robot during the interaction. Moreover, some results from the data analysis also suggest that the children were not very emotionally comfortable with the robot. Specifically, during moments related to the children's feelings it was shown that their eye gaze was less likely to be directed at the robot. The most likely explanation, as discussed in chapter 5, is for the child to make mental space for thinking. However, the relationship between the child and the robot could also be a factor. It could be an interesting direction of research to dig deeper into this relationship and how this relates to the child's gaze patterns. Subsequently, if

there is a correlation, it could be researched how robots can be altered in looks or behaviour to connect better emotionally with children.

Another related domain in which there is still progress to be made, is that of eye gaze tracking for children. A lot of progress has been made within the field of eye tracking in general (for adults), to the extent that eye tracking is now achievable with high accuracy and low costs (i.e. simple hardware and setups). Nevertheless, adults and thus adult behaviour is seen as the norm. Whereas head and eye positions might be relatively easily translated from adults' to children's shapes, the behaviour is unaccounted for. Through visual inspection of the video data it was found that many children tend to move around a lot on a chair, for example by frequently changing position or continuously rocking back and forth. In addition, given the presence of a screen or other object of interest, the children tend to move really close to this object in moments where they are relevant. As such, gaze calibration, i.e. the determination of a most likely head position and facial structure, is not very useful and might even limit the data quality compared to a highly flexible system. More research is needed to find the most optimal ways of dealing with this behaviour and to propose a more reliable algorithm.

Given the results of this study, one could train an STC model on all of the available data and implement an application that does live memorable moment detection, for example with a sliding window, and uses these results on-the-fly. The application can use the identification of memorable moments to, for example, suggest to the child what they might find interesting, which should enhance the quality of the interactions and the interest or excitement of the child. Another example could be that the robot revises the less memorable moments at the end of an interaction in order to increase the overall memorability of the content discussed. With such a system and a control group, the usefulness, correctness and efficiency of memorable moments detection could be scientifically and practically verified.

Finally, in this study the suggestion is made that children's (selective) memory is an indicator of their preferences or interests, especially when these are studied or recorded on the long term, but it should be researched to what extent this claim is supported through further experiments. Control experiments should be done that record the children's self-identified preferences and compare these to the content of the memorable moments model. This should further solidify the added value of the incorporation of such a system in cHRI settings.

7

Conclusion

Robots can be deployed in classroom settings to aid teachers in the development of children's mental and physical health and resistance. In this study it was researched how the interactions between robot and child could be improved and how the usefulness of the interaction can be maximized by leveraging different modalities of user information. More specifically, it was researched how and to what extent eye gaze tracking can be exploited to identify moments that the child remembers from a child-robot interaction. During the interaction in the experiment, the robot presents the children with numerous moral dilemmas and 2 options as to how to respond to each dilemma, out of which the child has to choose one and then discuss the choice. After the cHRI, the children identify which of the discussed scenarios they remember, which is then linked to their eye gaze patterns during the cHRI.

Several dedicated time series machine learning models, i.e. ROCKET, DrCIF, TDE, STC and HIVE-COTE v2, were trained on the gaze data. All models performed significantly better than chance on distinguishing memorized moments from not memorized moments, where STC had the highest F1 score of 0.603. Models HIVE-COTE v2 and STC also performed significantly better than chance on distinguishing the memory quality of the memorized moments into one of the groups 'low quality', 'medium quality' or 'high quality', where STC had the highest F1 score of 0.352. Although the results are modest, it shows that there is meaning within the data and that the field of study is promising. In addition, the functioning of the models shows a promise of easy application of memorable moments detection in real-life systems. It is expected that, in future studies, better results can be achieved by minimizing noise in the data, through better instruction of the participants and increased robustness of the robot, as well as by the design of a handcrafted intelligent model, leveraging the new-found domain formation.

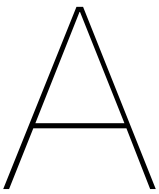
The data was also subject to manual pattern analyses, in which it was confirmed again that there is a meaningful relationship between memorability and gaze behaviour. It was found that children are significantly more likely to gaze at the robot during memorable moments. This is hypothesized to be due to a positive correlation between engagement, which has been linked to mutual gaze in literature, and memorability. It was also found that children are significantly less likely to gaze at the screen during memorable moments. This is hypothesized to be due to a positive correlation between cognitive processing, which has been linked to gaze aversion in literature, and memorability. Given the lack of other research into memorable detection, especially in the field of cHRI, these findings are substantial and give rise to more in-depth studies.

Another contribution of this study is the additional analysis into gaze patterns for the different levels of memory qualities and the reasons for remembering. In particular, it was shown that, during moments remembered with a lot detail, children are spending more time gazing at, or perhaps 'studying', the screen. Moreover, during moments experienced as cognitively easy by the children, it was shown that they look more at the robot and less at the screen. It is hypothesized that this is due to a lack of need for verification and comparison on screen, while engaging in the conversation with the robot. Finally, a significant difference was found between gaze patterns during moments remembered because of reasons related to feelings and those related to content, i.e. a notable item or person. The difference in gaze patterns follows expectations and children look more at the screen if they are interested in content and look more away to think about the moral dilemma.

The experiment and analyses conducted in this study are of an exploratory nature, in the context

of an unexplored field. Therefore, the fact that significant patterns were found should be an incentive for future research. It is believed that more significant results and/or more distinctive patterns can be found in a reproduction with a more organised and fail-proof experiment setup. In addition, the results of this study give rise to more research in related fields. In order to maximize a positive outcome of cHRIs, it is advised that more research is conducted on how to establish a social, emotional bond between robot and child. To achieve this, it is likely that better gaze tracking software needs to be developed that is robust in dealing with child behaviour. In addition, given the newly found knowledge in this study, special and dedicated intelligent models can be made that leverage gaze-specific data pattern information as well as information regarding the temporal structure of a discussion, on top of the data used in this study. Finally, from a societal standpoint, the results of this study suggest that memorable moments detection during cHRIs is feasible and this knowledge can be applied in real-life settings. This technique can be used, for example, to build user models of children's preferences over time and provide better and more personalised lessons in mental health awareness or other fields.

In conclusion, this study proves, for the first time, that children's gaze patterns during child-robot interactions are indicative of what they remember and also, at least to some extent, why and how well. It is up to future research to confirm, specify and apply this newfound knowledge.



Post-interaction Interview Instructions

A.1. Goals

- **Study goal:** leverage children's eye gaze information to identify a relevant moment during an interaction between a robot and child with visual aid.
- **Experiment goal:** collect eye gaze data and collect (self-identification of relevant moments.
- **Interview goal:** collect identification of relevant moments. The following are some points of attention:
  - A moment should be identified as relevant by the child.
  - A moment should be as specific as possible: a single point in time of limited duration.
  - Relating a moment to a visual cue (robot action or illustration) would be helpful and is desired.
  - The interviewer should not bias or steer the child's thought process.

A.2. Hypothesis

The visual focus of attention of the child, combined with eye movement patterns, fixations and saccades, is related to the child's memory activation.

A.3. Workflow

1. Child exits the room with the robot.
2. Escort the child to the next interview room. Preferably don't talk to the child at all, but especially not about their experience with the robot.
3. Sit the child down.
4. Start voice recording.
5. Execute script.
6. Stop voice recording.
7. Put the child to work on the paper questionnaire.
8. Check the validity of the recording. If necessary, fix any issues before the next session.
9. Before every big (> 5 min) break: make a backup of the data.
10. Make notes of unusual things or difficult items for the discussion.
11. Escort the child out of the room, back to the main activities.

A.4. Script

Interviewer (english)
Child
Notes

1. Hoi! Kan je me iets vertellen dat je je herinnert van je interview met de robot?
   Hi! Can you tell me something that you remember from the interview with the robot?
   If unsuccessful, repeat this question, phrased differently, e.g.
   - Als je later je ouders weer ziet, wat zou je ze vertellen over de ervaring die je net hebt gehad?
     When you see your parents later, what would you tell them about the experience you just had?
   - Kan je de dingen die je nog weet van je gesprek met de robot opnoemen?
     Can you list all the things you remember from your conversation with the robot?

   It is wise to take notes at this point, in case the child gives a lot of information. You can circle back to this later, to gather the 3 different moments.
2. [Vague answer, e.g.: the robot moved it's arms funny; I had to make decisions; I talked to a robot; I saw funny pictures]
3. Okee, weet je nog een specifiek moment dat dit voorkwam?
   Do you remember a specific moment that this happened?
4. Case:
   - [Yes: identifies some content-related moment in the conversation]
     (a) [repeat] Kan je je nog meer details herinneren van dit moment?
         Can you remember more details from this moment?
     (b) [No]
         i. Weet je misschien nog over welk scenario of welke keuze je hier nadacht?
            Do you perhaps remember the scenario/decision that you were thinking about at this time?
         ii. Wat weet je nog van wat er op het scherm stond tijdens deze gebeurtenis?
             Can you give more details about what was on the screen during this moment?
   - [No: it happened all the time / I don't remember]
     (a) [Try once more to ask for a specific moment related to this. If no success: return to point 1]
5. Waarom denk je dat dit moment je is bijgebleven?
   Why do you think that you remember this particular moment?
   Try to identify whether the 'why' is related to:
   - the robot's behaviour
   - the visuals on screen
   - the content / decision making process
6. [Free answer]
7. [Repeat until you have the minimum viable information for 3 moments.]
8. For the last 2 minutes OR when the conversation falls flat: Pull out copies of the scenario visuals that they were shown.
   Welke van deze afbeeldingen kan je je nog het beste herinneren? En waarom?
   Which of these images can you remember the best? And why?

B

Thematic analysis

labels	sublabels
fact_about_others	view
	social_facts
	unexpected_info
fact_about_world	entities
	people
self_perception	annotator_feelings
	annotator_stories
shared_experience	shared_story
meta_behaviour_of_other	emotional_moment
	behaviour
cognitive	cognitive_empathy
time_label	first
	last

Table B.1: Heuristics for for labelling humans' self-identified reasons for remembering (parts of) conversations as identified by Tsfasman et al., 2022.

Speaker	Text	Label	Memory quality
Interviewer	Kan je iets over één van de scenario's vertellen?		
Child	Dat was over het eten, dat ik, een meisje, het brood niet lekker vindt dat ze mee heeft. Dus wat zou jij doen, jouw eigen eten opeten wat je niet lekker vindt of ruilen als hij zijn eten niet lekker vindt. Ik zei "ruilen" omdat de ander dan nog wel wat te eten heeft en dan is iedereen blijer.	Indirect: empathy	2
Interviewer	En waarom denk je dat je je die goed herinnert?		
Child	Omdat ik van eten houd.	Direct: interesting content	

Table B.2: Excerpt + annotation of participant C24 talking about one of the scenarios discussed during the preceding cHRI.

Speaker	Text	Label	Memory quality
Child	Het beste kan ik me die herinneren.		
Interviewer	Scenario acht, en waarom?		
Child	Omdat het niet zo'n heel grote spraak was.	Direct: ability	
Interviewer	Ja, en wat deden ze daar dan?		
Child	Ze deed eerst... Ik dacht dat ze dacht... Nou, ik kan blijven zitten tot de hele klas niet gaat luisteren. Of ik steek mijn vinger er even op, dat ik het meteen kan zeggen tegen de klas. En ik dacht, vinger opsteken, dat dat is wel zo handig. Dan kan de juffrouw meteen weten, oh, zij wilt wat zeggen.	Indirect: empathy	2

Table B.3: Excerpt + annotation of participant C3 talking about one of the scenario's discussed during the preceding cHRI.

C

Exact performance of the machine learning models

C.1. Datasets

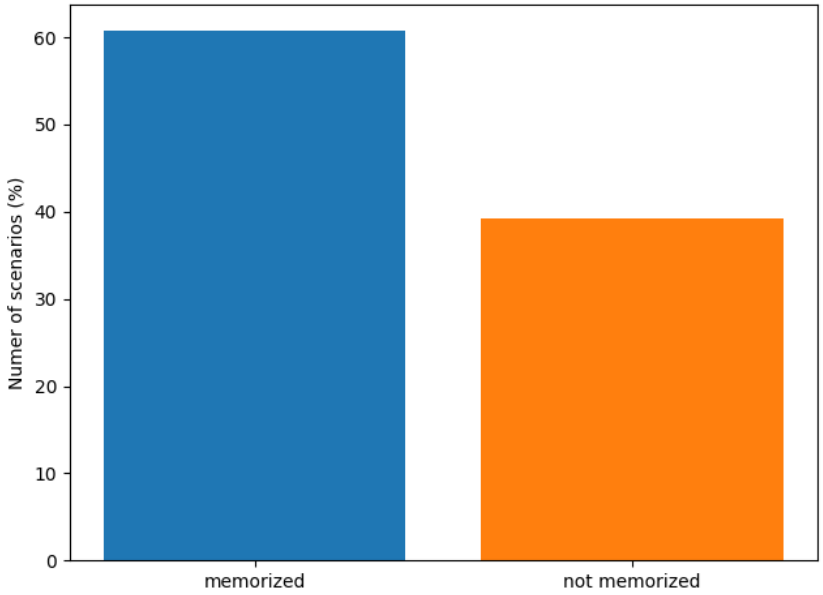


Figure C.1: The class distribution for the binary dataset.

C.2. Results in numbers

The models ROCKET, DrCIF, STC, TDE and HIVE-COTE v2 were trained on the eye gaze dataset in 4 distinct problem settings: the binary case (memorized vs not memorized samples), the memory quality case (samples labeled according to the quality of the memory descriptions), the reasons case (samples labeled according to the reason for remembering it) and the handpicked binary case (memorized vs not memorized samples of participants that behaved more compatibly with the gaze tracking algorithm). The models were trained and tested using 6-fold cross validation. As an evaluation metric, balanced accuracy and the F1-score are used. We define C to be the list of all classes in a dataset and the number of occurrences of a class $c \in C$ divided by the total number of classes is defined as w_c. TP,FP and FN stand for true positives, false positives and false negatives, respectively. Given these definitions,

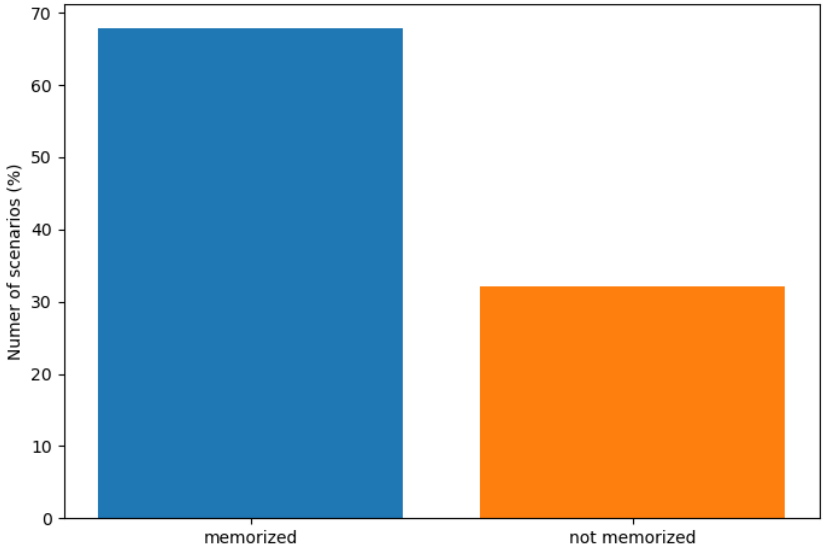


Figure C.2: The class distribution for the handpicked dataset.

balanced accuracy is defined as:

$$\sum_{c=1}^{C} w_c * \frac{TP_c}{TP_c + FN_c}$$

and the F1 score is defined as:

$$\sum_{c=1}^{C} w_c * \frac{TP_c}{TP_c + \frac{1}{2}(FP_c + FN_c)}$$

As such, the balanced accuracy is a weighted average of how many items in each actual class were correctly classified as this class. The weighted F1 score, on the other hand, is the weighted average of the harmonic mean between precision, $\frac{TP}{TP+FP}$, and recall, $\frac{TP}{TP+FN}$, and thus also takes into account the ratio of correctly classified cases over all cases predicted as that class. These scores for all problem/model combinations are listed in Table C.1.

C.3. Comparison per model

DrCIF Over all the problems, there was a significant difference in the F1 score between DrCIF ($M = .35, SD = .178$) and Dummy ($M = .24, SD = .105$); $t(48) = 2.619, p = .012$. There was not a significant difference in the F1 score for problem 'Reason dataset' between DrCIF ($M = .214, SD = .11$) and Dummy ($M = .194, SD = .1$); $t(12) = .34, p = .741$. There was a significant difference in the F1 score for problem 'Binary dataset' between DrCIF ($M = .209, SD = .07$) and Dummy ($M = .536, SD = .05$); $t(12) = 9.285, p < .001$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between DrCIF ($M = .322, SD = .116$) and Dummy ($M = .382, SD = .227$); $t(12) = .582, p = .573$. There was not a significant difference in the F1 score for problem 'Quality dataset' between DrCIF ($M = .213, SD = .099$) and Dummy ($M = .289, SD = .074$); $t(12) = 1.493, p = .166$.

ROCKET Over all the problems, there was a significant difference in the F1 score between ROCKET ($M = .368, SD = .159$) and Dummy ($M = .24, SD = .105$); $t(48) = 3.286, p = .002$. There was not a significant difference in the F1 score for problem 'Reason dataset' between ROCKET ($M = .214, SD = .11$) and Dummy ($M = .198, SD = .155$); $t(12) = .209, p = .839$. There was a significant difference in the F1 score for problem 'Binary dataset' between ROCKET ($M = .209, SD = .07$) and Dummy ($M = .516, SD = .062$); $t(12) = 8.007, p < .001$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between ROCKET ($M = .322, SD = .116$) and Dummy ($M = .436, SD = .106$); $t(12) = 1.775, p = .106$. There was not a significant difference in the F1 score for

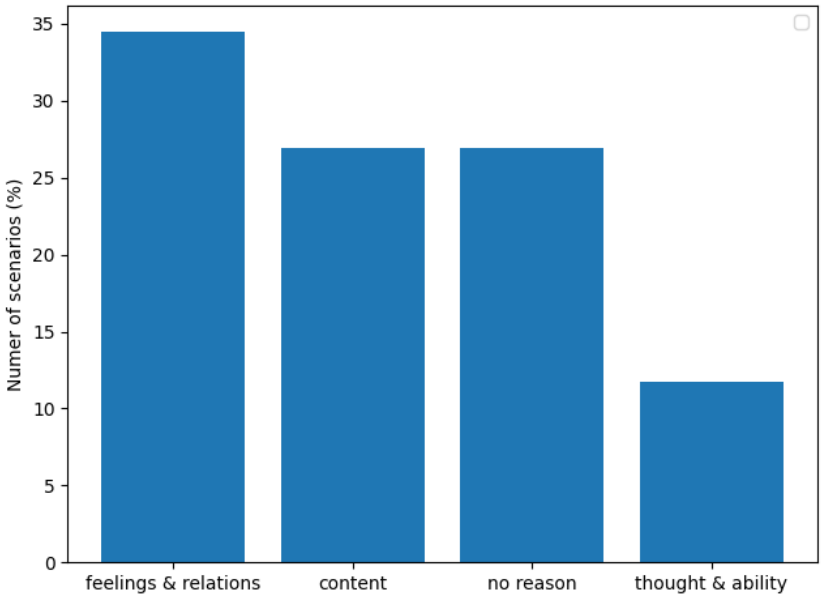


Figure C.3: The class distribution for the reason dataset.

problem 'Quality dataset' between ROCKET ($M = .213, SD = .099$) and Dummy ($M = .321, SD = .092$); $t(12) = 1.947, p = .08$.

STC Over all the problems, there was a significant difference in the F1 score between STC ($M = .422, SD = .205$) and Dummy ($M = .24, SD = .105$); $t(48) = 3.874, p < .001$. There was not a significant difference in the F1 score for problem 'Reason dataset' between STC ($M = .214, SD = .11$) and Dummy ($M = .263, SD = .068$); $t(12) = .92, p = .379$. There was a significant difference in the F1 score for problem 'Binary dataset' between STC ($M = .209, SD = .07$) and Dummy ($M = .603, SD = .087$); $t(12) = 8.633, p < .001$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between STC ($M = .322, SD = .116$) and Dummy ($M = .468, SD = .314$); $t(12) = 1.073, p = .308$. There was a significant difference in the F1 score for problem 'Quality dataset' between STC ($M = .213, SD = .099$) and Dummy ($M = .352, SD = .062$); $t(12) = 2.904, p = .016$.

TDE Over all the problems, there was a significant difference in the F1 score between TDE ($M = .34, SD = .186$) and Dummy ($M = .24, SD = .105$); $t(48) = 2.309, p = .025$. There was not a significant difference in the F1 score for problem 'Reason dataset' between TDE ($M = .214, SD = .11$) and Dummy ($M = .164, SD = .055$); $t(12) = 1.012, p = .335$. There was a significant difference in the F1 score for problem 'Binary dataset' between TDE ($M = .209, SD = .07$) and Dummy ($M = .528, SD = .096$); $t(12) = 6.565, p < .001$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between TDE ($M = .322, SD = .116$) and Dummy ($M = .397, SD = .214$); $t(12) = .755, p = .468$. There was not a significant difference in the F1 score for problem 'Quality dataset' between TDE ($M = .213, SD = .099$) and Dummy ($M = .273, SD = .112$); $t(12) = .976, p = .352$.

HIVE-COTE v2 Over all the problems, there was a significant difference in the F1 score between HIVE-COTE v2 ($M = .389, SD = .137$) and Dummy ($M = .24, SD = .105$); $t(48) = 4.256, p < .001$. There was not a significant difference in the F1 score for problem 'Reason dataset' between HIVE-COTE v2 ($M = .214, SD = .11$) and Dummy ($M = .237, SD = .059$); $t(12) = .437, p = .671$. There was a significant difference in the F1 score for problem 'Binary dataset' between HIVE-COTE v2 ($M = .209, SD = .07$) and Dummy ($M = .508, SD = .06$); $t(12) = 7.935, p < .001$. There was not a significant difference in the F1 score for problem 'Handpicked binary dataset' between HIVE-COTE v2 ($M = .322, SD = .116$) and Dummy ($M = .476, SD = .133$); $t(12) = 2.135, p = .059$. There was a significant difference in the F1 score for problem 'Quality dataset' between HIVE-COTE v2 ($M = .213, SD = .099$) and Dummy ($M = .337, SD = .062$); $t(12) = 2.594, p = .027$.

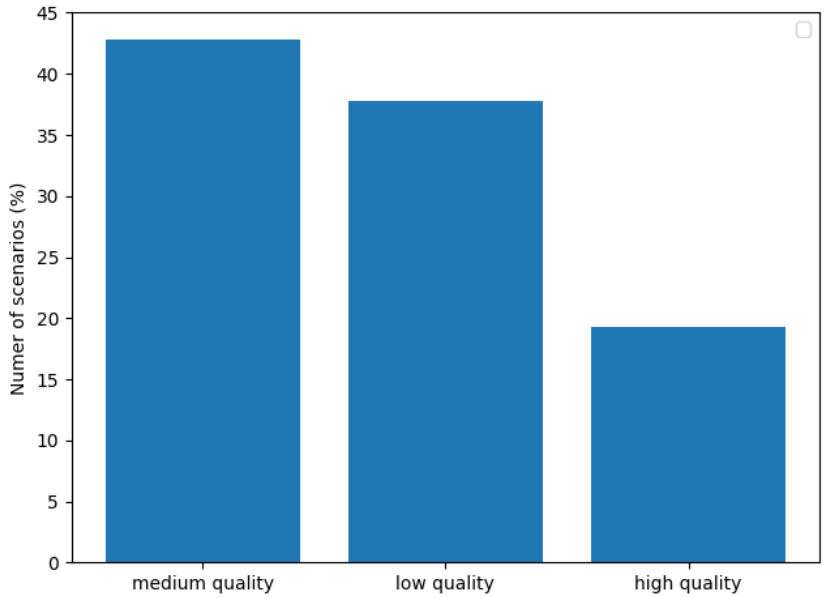


Figure C.4: The class distribution for the quality dataset.

problem	model	balanced accuracy	f1 score
Reason dataset	Dummy	0.208	0.214
Reason dataset	DrCIF	0.188	0.194
Reason dataset	ROCKET	0.232	0.198
Reason dataset	STC	0.233	0.263
Reason dataset	TDE	0.201	0.164
Reason dataset	HIVE-COTE v2	0.278	0.237
Binary dataset	Dummy	0.500	0.209
Binary dataset	DrCIF	0.535	0.536
Binary dataset	ROCKET	0.503	0.516
Binary dataset	STC	0.594	0.603
Binary dataset	TDE	0.498	0.528
Binary dataset	HIVE-COTE v2	0.486	0.508
Handpicked binary dataset	Dummy	0.500	0.322
Handpicked binary dataset	DrCIF	0.448	0.382
Handpicked binary dataset	ROCKET	0.482	0.436
Handpicked binary dataset	STC	0.524	0.468
Handpicked binary dataset	TDE	0.396	0.397
Handpicked Binary dataset	HIVE-COTE v2	0.446	0.476
Quality dataset	Dummy	0.250	0.213
Quality dataset	DrCIF	0.306	0.289
Quality dataset	ROCKET	0.311	0.321
Quality dataset	STC	0.388	0.352
Quality dataset	TDE	0.331	0.273
Quality dataset	HIVE-COTE v2	0.372	0.337

Table C.1: The classification results for the different machine learning models and the 4 different problem settings/datasets, as obtained with 6-fold cross validation.

Bibliography

Abbasi, N. I., Spitale, M., Anderson, J., Ford, T., Jones, P. B., & Gunes, H. (2022). Can Robots Help in the Evaluation of Mental Wellbeing in Children? An Empirical Study [ISSN: 1944-9437]. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1459–1466. https://doi.org/10.1109/RO-MAN53752.2022.9900843

Admoni, H., & Scassellati, B. (2017). Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, *6*(1), 25. https://doi.org/10.5898/JHRI.6.1.Admoni

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction [Series Title: Lecture Notes in Computer Science]. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, & V. C. Müller (Eds.), *Cognitive Behavioural Systems* (pp. 114–130). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34584-5_9

Andrist, S., Mutlu, B., & Gleicher, M. (2013). Conversational Gaze Aversion for Virtual Agents. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent Virtual Agents* (pp. 249–262). Springer. https://doi.org/10.1007/978-3-642-40415-3_22

Baddeley, A. (2013). *Essentials of Human Memory (Classic Edition)* [Google-Books-ID: 2YY3AAAAQBAJ]. Psychology Press.

Bagnall, A., Flynn, M., Large, J., Line, J., Bostrom, A., & Cawley, G. (2020). Is rotation forest the best classifier for problems with continuous features? [arXiv:1809.06705 [cs, stat]]. Retrieved May 15, 2023, from http://arxiv.org/abs/1809.06705

Bagnall, A., Bostrom, A., Large, J., & Lines, J. (2016). The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms. Extended Version.

Bagnall, A., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., & Keogh, E. (2018). The UEA multivariate time series classification archive, 2018 [arXiv:1811.00075 [cs, stat]]. https://doi.org/10.48550/arXiv.1811.00075

Bagnall, A., Flynn, M., Large, J., Lines, J., & Middlehurst, M. (2020). A tale of two toolkits, report the third: On the usage and performance of HIVE-COTE v1.0 [arXiv:2004.06069 [cs, stat]]. https://doi.org/10.1007/978-3-030-65742-0_1

Barragán-Sánchez, R., Romero-Tena, R., & García-López, M. (2023). Educational Robotics to Address Behavioral Problems in Early Childhood [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Education Sciences*, *13*(1), 22. https://doi.org/10.3390/educsci13010022

Baxter, P., Ashurst, E., Read, R., Kennedy, J., & Belpaeme, T. (2017). Robot education peers in a situated primary school study: Personalisation promotes child learning (N. Reich-Stiebert, Ed.). *PLOS ONE*, *12*(5), e0178126. https://doi.org/10.1371/journal.pone.0178126

Beattie, G. (2003). *: The New Psychology of Body Language*. Routledge. https://doi.org/10.4324/9780203500026

Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M., & Lim, A. (2017). UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 464–472. https://doi.org/10.1145/3136755.3136814

Bixler, R., & D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, *26*(1), 33–68. https://doi.org/10.1007/s11257-015-9167-1

Blythe, H. I., Liversedge, S. P., Joseph, H. S. S. L., White, S. J., & Rayner, K. (2009). Visual information capture during fixations in reading for children and adults. *Vision Research*, *49*(12), 1583–1591. https://doi.org/10.1016/j.visres.2009.03.015

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology [Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa]. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Breil, C., & Böckler, A. (2021). Look away to listen: The interplay of emotional context and eye contact in video conversations [Publisher: Routledge _eprint: https://doi.org/10.1080/13506285.2021.1908470]. *Visual Cognition*, *29*(5), 277–287. https://doi.org/10.1080/13506285.2021.1908470

Burdett, E. R. R., Ikari, S., & Nakawake, Y. (2022). British Children's and Adults' Perceptions of Robots (Z. Yan, Ed.). *Human Behavior and Emerging Technologies*, *2022*, 1–16. https://doi.org/10.1155/2022/3813820

Burgoon, J. K., Coker, D. A., & Coker, R. A. (1986). COMMUNICATIVE EFFECTS OF GAZE BEHAVIOR.: A Test of Two Contrasting Explanations. *Human Communication Research*, *12*(4), 495–524. https://doi.org/10.1111/j.1468-2958.1986.tb00089.x

Campos, J., Kennedy, J., & Lehman, J. F. (2018). Challenges in Exploiting Conversational Memory in Human-Agent Interaction. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1649–1657.

Cappella, E., Frazier, S. L., Atkins, M. S., Schoenwald, S. K., & Glisson, C. (2008). Enhancing Schools' Capacity to Support Children in Poverty: An Ecological Model of School-Based Mental Health Services. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*(5), 395–409. https://doi.org/10.1007/s10488-008-0182-y

Cegala, D. J., Sokuvitz, S., & Alexander, A. F. (1979). AN INVESTIGATION OF EYE GAZE AND ITS RELATION TO SELECTED VERBAL BEHAVIOR. *Human Communication Research*, *5*(2), 99–108. https://doi.org/10.1111/j.1468-2958.1979.tb00625.x

Coninx, A., Baxter, P., Oleari, E., Bellini, S., Bierman, B., Henkemans, O., Canamero, L., Cosi, P., Enescu, V., Espinoza, R., Hiolle, A., Humbert, R., Kiefer, B., Kruijff-Korbayova, I., Looije, R., Mosconi, M., Neerincx, M., Paci, G., Patsis, G., … Belpaeme, T. (2016). Towards Long-Term Social Child-Robot Interaction: Using Multi-Activity Switching to Engage Young Users [Accepted: 2017-06-27T12:59:02Z]. Retrieved January 25, 2023, from http://uhra.herts.ac.uk/handle/2299/18574

Coronado, E., Kiyokawa, T., Ricardez, G. A. G., Ramirez-Alpizar, I. G., Venture, G., & Yamanobe, N. (2022). Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems*, *63*, 392–410. https://doi.org/10.1016/j.jmsy.2022.04.007

Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., & Wilcock, G. (2012). Multimodal conversational interaction with a humanoid robot. *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 667–672. https://doi.org/10.1109/CogInfoCom.2012.6421935

Dabbagh, N., & Kitsantas, A. (2012). Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education*, *15*(1), 3–8. https://doi.org/10.1016/j.iheduc.2011.06.002

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., & Keogh, E. (2019). The UCR time series archive [Conference Name: IEEE/CAA Journal of Automatica Sinica]. *IEEE/CAA Journal of Automatica Sinica*, *6*(6), 1293–1305. https://doi.org/10.1109/JAS.2019.1911747

Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels [arXiv:1910.13051 [cs, stat]]. *Data Mining and Knowledge Discovery*, *34*(5), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z

Dhariyal, B., Le Nguyen, T., Gsponer, S., & Ifrim, G. (2020). An Examination of the State-of-the-Art for Multivariate Time Series Classification. *2020 International Conference on Data Mining Workshops (ICDMW)*, 243–250. https://doi.org/10.1109/ICDMW51313.2020.00042

Dini, A., Murko, C., Yahyanejad, S., Augsdorfer, U., Hofbaur, M., & Paletta, L. (2017). Measurement and prediction of situation awareness in human-robot interaction based on a framework of probabilistic attention. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4354–4361. https://doi.org/10.1109/IROS.2017.8206301

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, *70*(5), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology*, *38*(3), 438–445. https://doi.org/10.1037/0012-1649.38.3.438

Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition*, *33*(4), 727–733. https://doi.org/10.3758/BF03195338

Dudzik, B., Hung, H., Neerincx, M., & Broekens, J. (2018). Artificial Empathic Memory: Enabling Media Technologies to Better Understand Subjective User Experience. *Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions*, 1–8. https://doi.org/10.1145/3267799.3267801

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Edwards, B. I., & Cheok, A. D. (2018). Why Not Robot Teachers: Artificial Intelligence for Addressing Teacher Shortage [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2018.1464286]. *Applied Artificial Intelligence*, *32*(4), 345–360. https://doi.org/10.1080/08839514.2018.1464286

Elvir, M., Gonzalez, A. J., Walls, C., & Wilder, B. (2017). Remembering a Conversation – A Conversational Memory Architecture for Embodied Conversational Agents [Publisher: De Gruyter]. *Journal of Intelligent Systems*, *26*(1), 1–21. https://doi.org/10.1515/jisys-2015-0094

ePartners4All. (2021). Retrieved June 26, 2023, from https://epartners4all.com/ePartners4All

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review [arXiv:1809.04356 [cs, stat]]. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for Time Series Classification [arXiv:1909.04939 [cs, stat]]. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962. https://doi.org/10.1007/s10618-020-00710-y

Flynn, M., Large, J., & Bagnall, T. (2019). The Contract Random Interval Spectral Ensemble (c-RISE): The Effect of Contracting a Classifier on Accuracy. In H. Pérez García, L. Sánchez González, M. Castejón Limas, H. Quintián Pardo, & E. Corchado Rodríguez (Eds.), *Hybrid Artificial Intelligent Systems* (pp. 381–392). Springer International Publishing. https://doi.org/10.1007/978-3-030-29859-3_33

Gathercole, S. E. (1998). The Development of Memory [Publisher: Cambridge University Press]. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(1), 3–27. https://doi.org/10.1017/S0021963097001753

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658. https://doi.org/10.3758/BF03211385

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., & Maisonnier, B. (2008). The NAO humanoid: A combination of performance and affordability [arXiv:0807.3223 [cs]]. Retrieved January 25, 2023, from http://arxiv.org/abs/0807.3223

Han, J., Campbell, N., Jokinen, K., & Wilcock, G. (2012). Investigating the use of Non-verbal Cues in Human-Robot Interaction with a Nao robot. *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 679–683. https://doi.org/10.1109/CogInfoCom.2012.6421937

Harper, B., & Milman, N. B. (2016). One-to-One Technology in K–12 Classrooms: A Review of the Literature From 2004 Through 2014. *Journal of Research on Technology in Education*, *48*(2), 129–142. https://doi.org/10.1080/15391523.2016.1146564

Hasselbring, T. S., & Glaser, C. H. W. (2000). Use of Computer Technology to Help Students with Special Needs [Publisher: Princeton University]. *The Future of Children*, *10*(2), 102–122. https://doi.org/10.2307/1602691

Hirayama, T., Dodane, J.-B., Kawashima, H., & Matsuyama, T. (2010). Estimates of User Interest Using Timing Structures between Proactive Content-Display Updates and Eye Movements. *IEICE Transactions on Information and Systems*, *E93-D*(6), 1470–1478. https://doi.org/10.1587/transinf.E93.D.1470

Hodson, H. (2014). The first family robot. *New Scientist*, *223*(2978), 21. https://doi.org/10.1016/S0262-4079(14)61389-0

Hong, A., Lunscher, N., Hu, T., Tsuboi, Y., Zhang, X., Franco dos Reis Alves, S., Nejat, G., & Benhabib, B. (2021). A Multimodal Emotional Human–Robot Interaction Architecture for Social Robots Engaged in Bidirectional Communication [Conference Name: IEEE Transactions on Cybernetics]. *IEEE Transactions on Cybernetics*, *51*(12), 5954–5968. https://doi.org/10.1109/TCYB.2020.2974688

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). *Gaze-Based Detection of Mind Wandering during Lecture Viewing* (tech. rep.) [Publication Title: International Educational Data Mining Society ERIC Number: ED596576]. International Educational Data Mining Society. Retrieved November 24, 2022, from https://eric.ed.gov/?id=ED596576

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, *29*(4), 821–867. https://doi.org/10.1007/s11257-019-09228-5

Hutt, S., Krasich, K., R. Brockmole, J., & K. D'Mello, S. (2021). Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445269

Jermann, P., & Nüssli, M.-A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1125–1134. https://doi.org/10.1145/2145204.2145371

Jokinen, K. (2009). Gaze and Gesture Activity in Communication. In C. Stephanidis (Ed.), *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments* (pp. 537–546). Springer. https://doi.org/10.1007/978-3-642-02710-9_60

Kang, S.-H., Gratch, J., Sidner, C., Artstein, R., Huang, L., & Morency, L.-P. (2012). Towards building a virtual counselor: Modeling nonverbal behavior during intimate self-disclosure. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, 63–70.

Karna-Lin, E., Pihlainen-Bednarik, K., Sutinen, E., & Virnes, M. (2006). Can Robots Teach? Preliminary Results on Educational Robotics in Special Education [ISSN: 2161-377X]. *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, 319–321. https://doi.org/10.1109/ICALT.2006.1652433

Kim, A., Han, J., Jung, Y., & Lee, K. (2013). The effects of familiarity and robot gesture on user acceptance of information [ISSN: 2167-2148]. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 159–160. https://doi.org/10.1109/HRI.2013.6483550

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Large, J., Bagnall, A., Malinowski, S., & Tavenard, R. (2018). From BOP to BOSS and Beyond: Time Series Classification with Dictionary Based Classifiers [arXiv:1809.06751 [cs, stat]]. Retrieved May 31, 2023, from http://arxiv.org/abs/1809.06751

Lemaignan, S., Garcia, F., Jacq, A., & Dillenbourg, P. (2016). From real-time attention assessment to "with-me-ness" in human-robot interaction. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 157–164. https://doi.org/10.1109/HRI.2016.7451747

Lines, J., Davis, L. M., Hills, J., & Bagnall, A. (2012). A shapelet transform for time series classification. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 289–297. https://doi.org/10.1145/2339530.2339579

Lines, J., Taylor, S., & Bagnall, A. (2016). HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 1041–1046. https://doi.org/10.1109/ICDM.2016.0133

Loftus, G. R. (1972). Eye fixations and recognition memory for pictures. *Cognitive Psychology*, *3*(4), 525–551. https://doi.org/10.1016/0010-0285(72)90021-7

Mackworth, N., & Bruner, J. (2009). How Adults and Children Search and Recognize Pictures. *Human Development*, *13*(3), 149–177. https://doi.org/10.1159/000270887

Majgaard, G. (2015). Humanoid Robots in the Classroom. *IADIS International Journal on WWW/Internet*, *13*(1), 72–86.

Martelo, A. B., & Villaronga, E. F. (2017). Child-Robot Interaction Studies: From Lessons Learned to Guidelines. Retrieved June 27, 2023, from https://research.utwente.nl/en/publications/child-robot-interaction-studies-from-lessons-learned-to-guideline

McLoughlin, C., & Lee, M. J. W. (2010). Personalised and self regulated learning in the Web 2.0 era: International exemplars of innovative pedagogy using social software [Number: 1]. *Australasian Journal of Educational Technology*, *26*(1). https://doi.org/10.14742/ajet.1100

Middlehurst, M., Large, J., & Bagnall, A. (2020). The Canonical Interval Forest (CIF) Classifier for Time Series Classification. *2020 IEEE International Conference on Big Data (Big Data)*, 188–195. https://doi.org/10.1109/BigData50022.2020.9378424

Middlehurst, M., Large, J., Cawley, G., & Bagnall, A. (2021). The Temporal Dictionary Ensemble (TDE) Classifier for Time Series Classification [arXiv:2105.03841 [cs]]. https://doi.org/10.1007/978-3-030-67658-2_38

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., & Bagnall, A. (2021). HIVE-COTE 2.0: A new meta ensemble for time series classification. *Machine Learning*, *110*(11), 3211–3243. https://doi.org/10.1007/s10994-021-06057-9

Müller, M. (Ed.). (2007). Dynamic Time Warping. In *Information Retrieval for Music and Motion* (pp. 69–84). Springer. https://doi.org/10.1007/978-3-540-74048-3_4

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. *Proceedings of the 15th international conference on Intelligent user interfaces*, 139–148. https://doi.org/10.1145/1719970.1719990

Normoyle, A., Badler, J. B., Fan, T., Badler, N. I., Cassol, V. J., & Musse, S. R. (2013). Evaluating perceived trust from procedurally animated gaze. *Proceedings of Motion on Games*, 141–148. https://doi.org/10.1145/2522628.2522630

Oertel, C., Włodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze Patterns in Turn-Taking.

Pandey, A. K., & Gelin, R. (2018). A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind [Conference Name: IEEE Robotics & Automation Magazine]. *IEEE Robotics & Automation Magazine*, *25*(3), 40–48. https://doi.org/10.1109/MRA.2018.2833157

Paplu, S., Navarro, R. F., & Berns, K. (2022). Harnessing Long-term Memory for Personalized Human-Robot Interactions [ISSN: 2164-0580]. *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 377–382. https://doi.org/10.1109/Humanoids53995.2022.10000213

Ranglijst van het hoogste en laagste gemiddelde inkomen per inwoner van de gemeenten in Nederland (bijgewerkt 2023!) (2023). Retrieved June 28, 2023, from https://allecijfers.nl/ranglijst/hoogste-en-laagste-inkomen-per-gemeente-in-nederland/

Reese, E., & Brown, N. (2000). Reminiscing and recounting in the preschool years. *Applied Cognitive Psychology*, *14*(1), 1–17. https://doi.org/10.1002/(SICI)1099-0720(200001)14:1<1::AID-ACP625>3.0.CO;2-G

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye Movements During Mindless Reading [Publisher: SAGE Publications Inc]. *Psychological Science*, *21*(9), 1300–1310. https://doi.org/10.1177/0956797610378686

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers [Place: US Publisher: Educational Publishing Foundation]. *School Psychology Quarterly*, *26*(1), 1–13. https://doi.org/10.1037/a0022714

Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing Engagement in Human-Robot Interaction, 8.

Riedo, F., Rétornaz, P., Bergeron, L., Nyffeler, N., & Mondada, F. (2012). A Two Years Informal Learning Experience Using the Thymio Robot. In U. Rückert, S. Joaquin, & W. Felix (Eds.), *Advances in Autonomous Mini Robots* (pp. 37–48). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27482-4_7

Rodríguez, J. J., Alonso, C. J., & Maestro, J. A. (2005). Support vector machines of interval-based features for time series classification. *Knowledge-Based Systems*, *18*(4), 171–178. https://doi.org/10.1016/j.knosys.2004.10.007

Roediger, H. L., & Guynn, M. J. (1996). Chapter 7 - Retrieval Processes. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 197–236). Academic Press. https://doi.org/10.1016/B978-012102570-0/50009-4

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

Saleiro, M., Carmo, B., Rodrigues, J. M. F., & du Buf, J. M. H. (2013). A Low-Cost Classroom-Oriented Educational Robotics System. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics* (pp. 74–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-02675-6_8

Saravanan, A., Tsfasman, M., Neerincx, M. A., & Oertel, C. (2022). Giving Social Robots a Conversational Memory for Motivational Experience Sharing [ISSN: 1944-9437]. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 985–992. https://doi.org/10.1109/RO-MAN53752.2022.9900677

Schwartz, S. H. (2006). Basic Human Values: An Overview.

Sekmen, A., & Challa, P. (2013). Assessment of adaptive human–robot interactions. *Knowledge-Based Systems*, *42*, 49–59. https://doi.org/10.1016/j.knosys.2013.01.003

Sharma, K., Jermann, P., Dillenbourg, P., Prieto, L. P., D'Angelo, S., Gergle, D., Schneider, B., Rau, M., Pardos, Z., & Rummel, N. (2017). CSCL and eye-tracking: 12th International Conference on Computer Supported Collaborative Learning - Making a Difference: Prioritizing Equity and Access in CSCL, CSCL 2017 (B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim, Eds.) [Publisher: International Society of the Learning Sciences (ISLS)]. *Making a Difference*, 727–734. Retrieved May 12, 2023, from http://www.scopus.com/inward/record.url?scp=85073356510&partnerID=8YFLogxK

Siegfried, R., & Odobez, J.-M. (2022). Robust Unsupervised Gaze Calibration Using Conversation and Manipulation Attention Priors. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *18*(1), 1–27. https://doi.org/10.1145/3472622

Stolzenberg, S. N., McWilliams, K., & Lyon, T. D. (2018). Children's conversational memory regarding a minor transgression and a subsequent interview [Place: US Publisher: American Psychological Association]. *Psychology, Public Policy, and Law*, *24*(3), 379–392. https://doi.org/10.1037/law0000176

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning [arXiv:1602.07261 [cs]]. https://doi.org/10.48550/arXiv.1602.07261

Thomas, N., & O'Kane, C. (1998). The ethics of participatory research with children [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1099-0860.1998.tb00090.x]. *Children & Society*, *12*(5), 336–348. https://doi.org/10.1111/j.1099-0860.1998.tb00090.x

Tsfasman, M., Fenech, K., Tarvirdians, M., Lorincz, A., Jonker, C., & Oertel, C. (2022). Towards creating a conversational memory for long-term meeting support: Predicting memorable moments in multi-party conversations through eye-gaze. *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 94–104. https://doi.org/10.1145/3536221.3556613

Wang, N., & Gratch, J. (2010). Don't just stare at me! *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1241–1250. https://doi.org/10.1145/1753326.1753513

Wang, Z., Yan, W., & Oates, T. (2016). Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline [arXiv:1611.06455 [cs, stat]]. Retrieved June 5, 2023, from http://arxiv.org/abs/1611.06455