Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Project No. 611391

DREAM
Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 611391

D5.3 Methods for improving the assessment

Due date: **M48**
Submission Date: **M60**

Start date of project: **01/04/2014** Duration: **54+6 months**

Organisation name of lead contractor for this deliverable: **University of Skövde**

Responsible Person: **Serge Thill** Revision: **2.0**

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary

This deliverable forms the final output of WP5. It reports both on the improvement analysis of the DREAM system performance, and on the continued work to develop algorithms that can infer internal states based on kinematics. Briefly, the deliverable covers

- A brief updated overview of the relevant state of the art, since this is the final time this work will be documented as part of the DREAM project
- Analysis of the performance of the DREAM system, including both a qualitative analysis of challenging cases for assessment, and a updated algorithm for turn-taking performance assessment
- A new study to assert that humans have access to relevant internal states from child movements – part of the PinSoRo dataset – in light of previous difficulties to obtain reasonable results from algorithmic approaches
- Results that demonstrate conceptors (a conceptual relative of Echo State Networks that we previously used) are in principle capable of this classification task as well, beyond what our previous attempts have producced
- Ongoing work based on a related approach, using the neuro-engineering framework, that promises to outperform the conceptor approach.

Principal Contributors

The main authors of this deliverable are as follows (in alphabetical order).

Madeleine Bartlett, University of Plymouth
Erik Billing, University of Skövde
Daniel Hernandez-Garcia, University of Plymouth
Vipul Nair, University of Skövde
Serge Thill, University of Skövde

Revision History

Version 2.0 (E.B, M.B. 09-04-2019)
New RC with updated section 3 and checks of section 2, 4, 5, and 6.

Version 1.0 (E.B, S.T. 08-04-2019)
First RC.

Version 0.1 (S.T. 21-03-2019)
Initial document

1 Introduction

The main purpose of this deliverable is to conclude the reporting on the work carried out as part of WP5 (and as such, its delivery has been delayed to coincide with the new end date of DREAM, following the extension). This includes its main purpose of reporting on the assessment methods and identifying possible improvements, and also the additional purpose requested by the previous deliverable 5.4 to report on improvement in the more general recognition of internal states from kinematics.

Since this is the final deliverable for this work as part of the DREAM work package, it is worth noting that the second aspect has broader relevance in the field of HRI (which we discuss further below). For that reason, we also include a brief discussion of the current state of the art that is kept somewhat general in section 2. Section 3 is then dedicated to reporting improvements in the performance assessment in the DREAM system. Since none of the previous approaches to inferring internal states from observed kinematics (reported in the previous deliverables) have proven to be particularly satisfactory, section 4 presents a small study to re-assert the degree to which human observers can perform this task.

Section 5 then presents a novel algorithmic approach, based on conceptors (which are conceptually related to the Echo State Networks previously used in an attempt to solve this problem), to inferring intentions from observations, which outperforms our previous attempts. Finally, section 6 is dedicated to ongoing work, which will continue beyond the end of DREAM. This includes, in particular, another algorithmic approach based on a delay network implemented in the Neuro-Engineering Framework (Eliasmith and Anderson, 2003), which has only very recently (Voelker and Eliasmith, 2018) been shown to optimally solve tasks that Echo State Networks address heuristically, and therefore has potential to outperform the conceptor-based approach.

2 Brief overview of the state of the art

This section presents different theories on how humans perceive and interpret the internal states of others, as well as an overview of previous research on how to enable computational systems and robots to mimic some of these functions. We conclude by highlighting the shortcomings of current techniques in dealing with a wide range of internal states, and thereby the aspects we develop further in this deliverable.

2.1 Human Mind-Reading

"Mind-reading", also referred to as mentalizing or theory of mind, is the human ability to infer the mental states of others (Premack and Woodruff, 1978). A large amount of research has explored how humans achieve this insight. The resultant theories and knowledge of the possible mechanisms behind this ability are central to the question of designing artificial systems with similar capabilities. Importantly, many, if not all, of these theories posit that humans use observable behavioural cues as indicators of internal states. The first step here is to identify the observable behaviours which humans use to communicate and recognize their internal states, and which could therefore be used by artificial systems.

Studies examining the mirror neuron system (MNS) found in primates and humans indicate that humans use observed kinematics or body movements to make inferences about the observed actor (Iacoboni et al., 2005; Gallese et al., 1996). Theories pertaining to the MNS propose that recognition is a result of an observer mapping the observed kinematics onto their own motor system, allowing them to simulate a representation of the intentions driving the observed action (Gallese et al., 1996).

Even outside of MNS research, a large body of research indicates that humans use observable cues, such as action kinematics (Lewkowicz et al., 2013) and facial expressions (Ekman and Friesen, 1971), to interpret the internal states of others. Given that action kinematics are a readily available resource for artificial systems, this seems a promising data source for designing an internal state classifier.

The majority of evidence supporting the theory that humans use observable biological movement to make internal state inferences comes from studies using stimuli in the style of point-light displays. These stimuli isolate information about an actors movements by presenting dots representing the actors joints, or stick-figures, moving against a blank background. This allows researchers to examine what information humans can recognize from only movement information. For example, Clarke et al. (2005) created their stimuli by filming pairs of actors performing a dialogue whilst portraying a particular emotion (e.g. fear, disgust, joy). Using point-light versions of these videos as stimuli Clarke et al. (2005) found that participants were able to recognize the portrayed emotional state from the biological motion alone. Similarly, Manera et al. (2011) showed participants point-light videos of actors performing a reach-to-grasp action. This action was motivated by one of 3 socially-relevant intentions: (1) cooperation, (2) competition, or (3) performing an individual action. Manera et al. (2011) found that participants were able to predict the goal, i.e. the social intention, of the action based on only the motion information. Such research suggests that internal state information is accessible in human movements. Becchio et al. (2017) outlined the "observability principle" to describe this idea, which posits that humans are able to directly perceive the internal states of others via differences in observable actions/movements.

The culmination of this evidence suggests that observable behaviours such as physical actions, body poses and facial expressions may be a useful source of data for a computational internal state recognition system. Examining this idea forms the basis for the study described in section 3.2.1.

2.2 Human Internal States

A second point of interest for this research project is the nature of the internal states we are attempting to classify. Theories and research from Psychology suggest that many internal states are described as continuous dimensions, rather than discrete categories (Russell, 1980; Eysenck, 1950). For example, whilst one could argue that a person is either certain or uncertain about how to perform a task, the degree of un/certainty may vary depending on the task.

When designing automated internal state classification systems it is important to consider what definition of the internal state is most appropriate. For example, if a collaborative robot is designed to ask "Do you need me to clarify the instructions?" whenever they notice their human partner is uncertain, then a binary classification may be sufficient. However, say a tutor robot is designed to either provide support to a child if they are struggling with a learning task, or to change to an easier task if the child is experiencing extreme difficulty, then the system for classifying "experienced difficulty" would need a more dynamic representation of experienced difficulty. That is, the identified state would need to fall on a continuous dimension ranging from "extreme difficulty" to "extreme ease" with several intermediate "levels".

Consequently, deciding the best definition for an internal state relies on a number of factors. This project is focused on developing a system able to provide detailed, human-like insight into recognized internal-states. As such, we will adhere, as closely as possible, to the definitions provided by Psychology. A secondary set of factors which we must consider, however, are the potential situations and tasks to which the classification systems we aim to develop could be applied. Below we discuss each of the settings we consider and how they and the involved human internal-states (as described in Psychology literature) will guide the development of classification techniques within this project.

2.2.1 Behaviour Classification for Autism Diagnosis

The Diagnostic and Statistical Manual of Mental Disorders (DSM-V) (American Psychiatric Association, 2013) defines Autism Spectrum Disorder (ASD) in terms of difficulties in two behavioural domains: social communication and interaction, and restricted or repetitive behaviours and interests (American Psychiatric Association, 2013). Importantly, The DSM-V and numerous other diagnostic tools emphasize the "spectrum" nature of ASD. This concept refers to: (1) differences in symptom presentation and severity within the clinical population, (2) the continuous distribution of "ASD-typical traits" between the general and clinical populations, and (3) subgroups (Lai et al., 2013). Diagnosis of ASD is, therefore, best thought of as a series of severity scores for each relevant symptom/behaviour. This describes the approach taken by a number of diagnostic tools designed to help clinicians in making diagnostic decisions. For example, the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000) provides clinicians with a set of prompts to elicit particular behaviours from individuals being assessed. The individual's responses to these prompts are then scored on a 4-point scale ranging from "no sign of ASD" to "severe ASD". Overall, the diagnosis of ASD relies on subjective interpretations by experts of observations of a child's behaviour made by clinicians, caregivers and teachers (Scassellati, 005a; Rogers et al., 2016). The subjective nature of diagnosis, and the high degree of clinical heterogeneity within ASD cases (Scassellati et al., 2012; Geschwind and Levitt, 2007), means that the process could be improved by introducing more quantitative, objective measures of child behaviour.

These benefits can be provided by introducing automated systems to aid in identifying the behaviours which indicate different levels of severity. Diagnostic systems to augment the diagnosis of ASD have been developed (Wall et al., 012a; Bone et al., 2015). However, in contrast with the diagnostic requirements, these systems usually approach behaviour classification in a binary fashion to provide a final diagnosis. For example, Wall and colleagues (Wall et al., 012a,b) designed an algorithm using machine learning methods which was able to differentiate between cases of individuals with and without ASD. Similarly Zunino et al. (2018) trained a Long-Short Term Memory (LSTM) model to classify children with and without ASD based on the kinematics of a reach-to-grasp motion. One limitation in both of these examples is that classification is binary. This lack of sensitivity to, or detail about, intermediate cases brings with it the ethical issues of overly simplified diagnostic measures, and of potentially classifying a large proportion of the behaviours which fall on the autism spectrum as non-ASD (Bone et al., 2015). Additionally, by providing the final diagnosis, these approaches assume the role of the clinician. There are a number of pitfalls associated with this, the main one being that it assumes the classification system is as good as, or better than, human clinicians and experts. An alternative approach would be to *support* the clinician's decision making process by providing detailed descriptions of the child's behaviour.

In order to provide richer, descriptive analyses of a child's behaviours, a system must possess several classification categories ranging from "typical of the general population" to "severely atypical". Whilst this can be achieved using classical machine learning techniques, such an approach would require a large training data-set which includes examples of each behaviour, at each level of severity. Such a data-set would be difficult and time-consuming to obtain. We therefore require a method able to deal with the spectrum nature of ASD. One way to do this would be to apply a method which can represent behaviours as points along a continuous dimension of severity, and which requires less data for learning than traditional machine learning methods. An initial study evaluating one approach in this direction is described in 3.2.2.

2.2.2 Robots for ASD Interventions

As well as arguing for the use of automated behaviour classification systems in ASD diagnosis, we propose that embedding these systems into socially interactive robots can provide further benefits during ASD interventions. Research suggests that individuals with ASD respond better to interactions with technology (Ozonoff, 1995) and are more likely to produce rare social behaviours in interactions with robots compared to with humans (Ricks and Colton, 2010). Consequently, work has been done to examine and improve the effectiveness of robots in clinical settings. The focus of much of this research has been the use of robots in interventions; to encourage the performance of social behaviours and mediate the transfer of these skills to interactions with other humans (Esteban et al., 2017; Duquette et al., 2008; De Silva et al., 2009; Robins et al., 2005; Scassellati et al., 2012; Thill et al., 2012).

On-line behaviour adaptation is important for autonomous robots in ASD interventions due to the high variability seen between children with ASD (Scassellati, 005b). This process requires the system to track and classify the child's behaviour before appropriate responses can be selected. Furthermore, recognizing any changes in a child's abilities or behaviours that occur as a result of the intervention provides an opportunity to have a robot autonomously adapt to these changes, adjusting the intervention task or level of difficulty, without overly relying on the clinician. This, in turn, would promote smoother, more fluent robot assisted interventions, reducing the potential for stress or confusion on the participant. Measuring a child's behaviour directly during intervention interactions would also allow robots involved in the intervention scenario to track a child's progress and provide clinicians with quantitative measures of how a child is improving.

2.2.3 Behaviour Classification for Tutor Robots

Outside of diagnostic situations, numerous human-robot interaction scenarios would benefit from having a robot able to recognize the internal states of their interaction partner. We focus on tutoring scenarios where a robot is employed to provide guidance and supervision in a one-to-one teaching interaction with a child. Whilst evidence suggests that such one-to-one interactions can promote learning in students (Cohen et al., 1982; Bloom, 1984), this approach is often not possible in classrooms where single teachers (in the U.K.) are responsible for 27 students on average (Department for Education, 2018).

To address this, research has been conducted into developing technologies, such as robots, able to provide more personalized teaching interactions. For example, Hood et al. (2015) developed a robot to help children improve their handwriting skills, and the L2TOR project has worked to develop a robot system for teaching children a second language (Belpaeme et al., 2015). Furthermore, research has demonstrated that robot tutors providing personalized interactions do lead to learning benefits. For example, Leyzberg et al. (2014) designed a robot to help participants solve puzzles by providing feedback. In one condition, the feedback the robot gave was selected based on an assessment of the participant's skills (personalized feedback), whilst in the other condition feedback was random (control). Leyzberg et al. (2014) found that participants who received personalized feedback showed greater improvements in a post-test than participants in the control condition. We argue that children's internal states are an additional metric which could be harnessed for providing personalized interactions in such settings. For instance, if a robot tutor is able to recognize a child's level of experienced difficulty with a task the robot could be provided with action policies to respond to these levels, e.g. providing encouragement and support for "medium" difficulty, and switching to an easier task for "high" difficulty. Importantly, we consider that the more "levels" of an internal state that a robot system is able to recognize, the more adaptive and fluent a robot's action policy can be.

2.3 Automated Internal State Recognition: The State of the Art

Whilst it is clear that different scenarios require different definitions of internal states for classification, an approach which uses a continuous representation for this task is yet to be developed. Here we review some of the available classification techniques which have been applied to this problem, and discuss one potential alternative approach.

Various methods have been developed for the on-line classification of human behaviour in terms of human internal states. One approach draws from a particular school of thought within Psychology which posits that humans are able to recognize the intentions of others based on observable movements and actions by mapping what they observe onto their own experiences or reasoning about what they would do in that particular situation (Gallese and Goldman, 1998). So, for example, one theory proposes that when we observe someone reaching for an object, we "mirror" the observed movement in the motor and motion planning regions of our brain and then draw on our experiences of performing similar movements to infer the intentions behind the observed action (Goldman et al., 2012). This school of thought has led HRI researchers to develop systems which recognize human internal states by using the system's own experience to map observed actions onto intentions. For example, Kelley et al. (2008) designed a robot to recognize actions and intentions using its own experience. Hidden Markov Models (HMMs) were used to model five activities and the robot was made to perform these activities to train the HMMs. By monitoring the differences in features of its own performance for each action the robot system produced a mapping between observable action features and internal intentions. As a result the robot was able to correctly identify which activity an observed human was performing. Despite this success, the reliance on the robot's own experience limits this approach due to the need for a to-be-recognized activity or internal state to be added to the robot's behavioural repertoire. As there are a number of internal states which robots cannot experience, such as emotional states, this method is limited to the recognition of action intentions.

Alternative approaches do not rely on a robot's own experiences. However, many of these approaches are fairly limited in terms of the number of internal states they're able to recognize. For example, Franois et al. (2007) developed a system using Self-Organizing Maps, which would allow a robot to classify the style in which a child played with it as either "gentle" or "strong". The robot used this classification to select appropriate responses to the child's behaviour. Whilst this approach was successful in interactions with children, it requires a very specific scenario with limited interaction styles and response options. Additionally, this approach relied on detecting the nature of haptic interactions with the robot. In accordance with the discussed theories of how humans may recognize the internal states of others, other approaches have used observable behaviours as the input for estimating a human's internal state. Daoudi et al. (May), for example, designed a system for distinguishing between social and personal intentions based on the observable characteristics of a human's reach-grasp-lift-place action. Principle Component Analysis (PCA) was used to define each intention category ("social" and "personal"). The resulting system demonstrated a high success rate (68%). Together, these studies illustrate that, whilst existing approaches to internal state recognition are successfully able to recognize internal states from observable behaviours, they are limited to identifying a small number of internal states.

Even approaches which incorporate a less limited number of classification labels, still use a fairly rigid, categorical characterization of internal states. To illustrate, Wimmer et al. (2008) developed a system for recognizing human emotional states based on facial expressions. Wimmer et al. used a Binary Decision Tree as the classifier and trained and tested it using the Cohn-Kanade Facial Expression Database (Kanade et al., 2000) which consists of images of people showing the six universal facial expressions defined by Ekman and Friesen (1971). The resultant system performed well on

identifying the different emotions expressed in the still images. However, this approach results in a description of emotional states such that a person is either happy or not happy. Whilst this is sufficient for some situations where robots would benefit from being able to recognize emotions, there are scenarios where more detail is needed to provide more appropriate robot responses and behavioural protocols. For example, say we have a robot situated in a classroom which is required, among other things, to identify conflicts between children and to decide whether to intervene and mediate the interaction, or to attract the attention of an adult human to intervene. If that robot is only able to identify whether or not a child is angry, and not *how* angry, the possible responses available to that system are limited and would result in inappropriate behaviour from the robot. This is the same for other internal states including task engagement and experienced difficulty.

2.4 Summary

This discussion demonstrates that the current state-of-the-art for systems designed to recognize the internal states of humans is subject to a number of limitations. There are two main limitations of key interest to this project. The first is the restricted number of categories such systems can practically be trained to recognize. This limitation is usually a result of the amount of training data required to achieve accurate classification. For instance, the system designed by Daoudi et al. (May) was trained on 375 examples of actions in each category. Scaling a system up to include more categories using this type of approach would require an impractically extensive training dataset. Consequently, one goal for developing a more comprehensive internal state recognition system is to design a system which can be trained on a relatively small data set.

The second limitation is the restrictive nature of a categorical approach to internal states. That is, classification systems such as the one developed by Wimmer et al. (2008) consider emotions as discrete states whereby a person either does or does not experience the emotional state. For robots to respond more appropriately to human internal states, recognition systems should provide more detailed representations of these internal states such that they reflect the human experience more accurately. As discussed above, numerous internal states are thought to be best described as continuous dimensions rather than discrete states. So for a robot to appropriately respond to a human's experienced *level* of an internal state, e.g. the optimal responses to *high* and *low* levels of frustration may differ in a situation, the recognition system must provide a different classification for each "level".

Overall, this therefore identifies three requirements for an internal state classification system:

1. Internal state information must be available in the chosen data source
2. Internal states should be represented as points along continuous dimensions
3. Classification systems should be trainable with limited datasets

One approach which we believe may achieve this is to use conceptors (Jaeger, 2017). Conceptors are defined as a neuro-computational mechanisms that can be used for learning a large number of dynamical patterns. The main benefits of this approach to our goals are that they can be trained using only examples of the extremes along a continuous dimension and then combined to generate the intermediate patterns. Additionally, this approach assumes that underlying any behaviour is a continuous dimension, which suits our requirement for representing multiple "levels" or "intensities" of a given internal state. We describe work in this direction in section 5

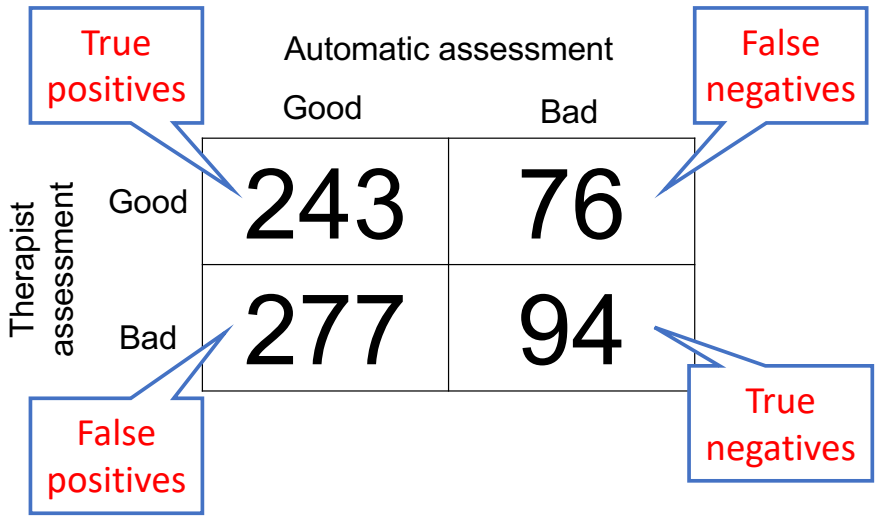

Figure 1: System assessment performance as reported at the P3 review. The agreement between human and system ratings is 49%.

3 Improving the performance assessment in the DREAM system

One of the main tasks in WP5 has been to develop methods for assessing the child's performance during robot enhanced therapy (RET). This concerns performance assessment during three different tasks, *turn-taking (TT), joint attention (JA)*, and *imitation (IM)* (D1.1). The performance measure varies depending on the task.

During TT, performance is defined as the child's ability to wait for its turn during a turn-based game that the child and the interaction partner (robot in RET condition, human in SHT condition) play on the sand tray (i.e., the large touch screen located between child and the robot). Specifically, the child is expected not to move over the sand-tray during the opponent's turn. This duration is referrer to as the *assessment window*. The beginning of the assessment window is indicated by the robot saying *Now is my turn (Acum e rndul meu)*, ends with a *"Pling"* sound, and lasts for about 8.5 seconds. During this time, participants are instructed not to move over the sand tray.

For JA and IM tasks, the performance measure was identical to the performance on task. That is, in the JA condition, the child should follow the robot's gaze and look at one out of two target images. In the IM condition, the child should imitate the robot's behavior. Please refer to D1.1 for details.

From a system point of view, this functionality was implemented as the *assessChildPerformance* component. As reported during the P3 review, the system performance, measured as its ability to produce similar performance assessments as a human therapist, was far from satisfactory, (c.f., Figure 1). Primarily, the system produced a large number of False negatives, that is, cases where the system assesses bad performance while the human therapist classifies the child's performance as good.

While we initially thought that this poor performance was primarily due to system calibration and tuning issues, we later realized that this was not the case. Although the performance measures in DREAM were well specified for all DREAM intervention tasks, a human annotator always adapts the

assessment based on factors that were not known at design time. An analysis with examples of cases where the initial performance definition has to be updated is presented in Section 3.1.

In order to better understand the challenges behind the automatic assessment, a deeper analysis was made, presented in Section 3.1. This analysis lead to a re-design of the assessment methods, initiated early 2018. At this point, a majority of the third and final evaluation of the full DREAM system was already completed. In order to not interfere with clinical results of the evaluation, we made the decision not to implement any dramatic changes to the live system. Instead, the *assessChildPerformance* component was re-implemented to allow offline assessment of recorded interventions. In addition to increased assessment performance, an important contribution of this work is a fully traceable assessment process where the video annotation tool ELAN[1] is used as a data visualization tool between therapists and algorithm designers. An key component of this work is *sensoryAnalysisOffline* that allows re-playing of sensor perceptions from recorded data. The algorithms behind *sensoryAnalysisOffline* is presented in detail in Cai et al. (2018) with a complete route to traceable assessments presented in D3.4.5. The updated methods for performance assessment, focusing on turn-taking, are presented in Section 3.2. This work was also complemented with a deep-learning/conceptor based approach to assessing internal state information from observable clues more broadly. This part of the work is presented in sections 4 and 5.

3.1 Challenges in performance assessment

With the goal of providing a better understanding of the challenges involved in performance assessment, we here present a few of typical examples where therapists and the automatic assessment system disagree. As mentioned in the previous section, the system made a large proportion of false negatives and thus, we've focused on these situations. There were frequent errors in all three task types (TT, JA, and IM), but the majority came from turn-taking and therefore this analysis focuses on these cases.

Despite the fact that we started out with fairly simple and manually designed assessment algorithms, it is in many situations very difficult to see exactly why the system produces a specific assessment merely by inspection of the video. In this work, the automatically generated ELAN files, with annotations indicating the system's assessment variables, constituted a critical tool for analysis and discussions between therapists and algorithm designers.

In the turn-taking tasks, the largest source of false negatives, i.e., situations where the system assesses a "bad performance" on a task which the therapists assess as "good performance", comes from *visual stimulation*. Many children in the study frequently looks closely at the sand tray during different tasks. This behavior is typical among children diagnosed with ASD and is seen as a type of observation, and is thus not regarded as an interference with the opponent's turn. An example of this behavior is visible in Figure 2. Although accepted by the therapists, it constitutes a big motion over the sand tray that is otherwise not allowed. Additionally, the hands are frequently hidden under the child's chest, making it difficult for the system to reliably determine the hand position and thus, the waiting performance.

Another common reason for false negatives were hidden hands, frequently appearing when the child is leaning over the sand tray. An example of this is visible in Figure 3. In these cases, the skeleton pose calculation makes an incorrect estimation of hand position, indicating that the hands are in front of the child. As a result, the system sometimes assesses these situations as "bad waiting" although the behavior is compliant with instructions.

A third source of false negatives is presented in Figure 4. As indicated in the lower part of the

[1]ELAN Annotation Tool by the Max Planck Institute: `https://tla.mpi.nl/tools/tla-tools/elan/`

Figure 2: An example of *visual stimulation*. The child is looking closely at the sand tray. It is regarded an acceptable behavior during the opponent's turn since it constitute a form a looking.

figure, we are approaching the end of the robot's turn. The system is correctly estimating the location of both arms, one held close to the chest and the other over the sand tray. Looking only at this image, it constitutes a clear example of a child not able to wait. One hand is clearly over the sand tray, which is not allowed during the robot's turn.

However, when observing the context of this action, the interpretation changes. The sequence of events leading up to Figure 4 is presented in Figure 5. During this period, the child is following the robot's move by looking at the object mooving at the screen. He is waiting for the object to reach its target location, but as soon as it has, he moves the arm out over the sand tray. His actions are clearly coordinated with the robot and his intention is not to interfere, thus the behavior is assessed as good waiting.

In sum, the analysis displays a range of challenges in performance assessment. Some of these are due to technical limitations in e.g., pose estimation of the hands, but the majority of wrong assessments are not caused by perceptual limitations. Instead, the challenge lies in capturing the child's *intention* behind different actions. The intention of in child in Figure 2 is to look, not to engage in the robot's turn. Likewise, the intention of the child in Figure 4 is to wait, altough he did not exactly follow the definition of a turn as specified at design time.

3.2 Updated methods for performance assessment

Based on the analysis presented in the previous sections, the methods for performance assessment was re-implemented for offline analysis. Here, we focused on assessment of waiting performance during turn-taking. Joint attention and imitation performance required significantly more work and also constituted a smaller proportion of the complete dataset, and were therefore not included in the

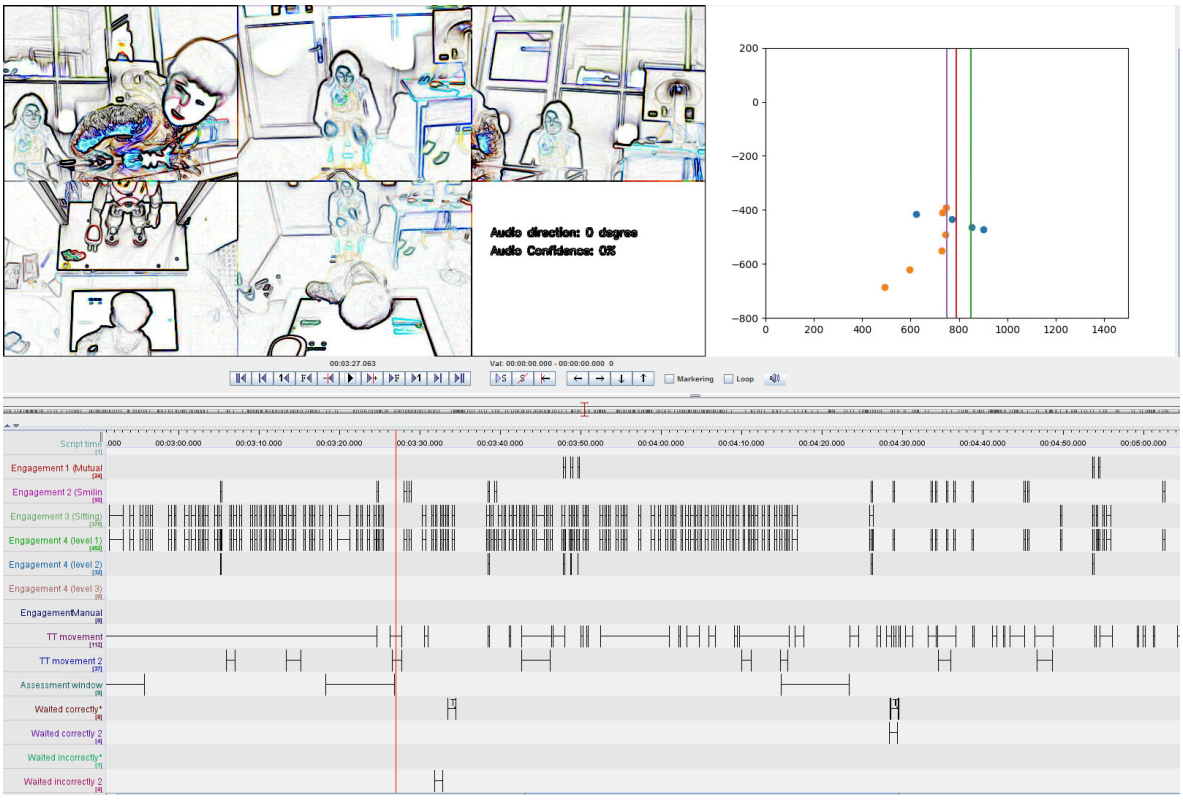

Figure 3: A screen shoot from the ELAN annotation tool. The child is following the movement of the animated car during the robot's turn. Since the child's hands are hidden and the child is leaning over the sand tray, it constitutes a challenging situation for the skeleton calculation. The right diagram presents the estimated joint positions seen from the right. The hands, marked by orange dots, are incorrectly positioned in front of the child, causing the system to assign "bad waiting" although the child is in fact following instructions, keeping his hands away. The lower part of the figure displays automatically generated annotations indicating different aspects of system information.

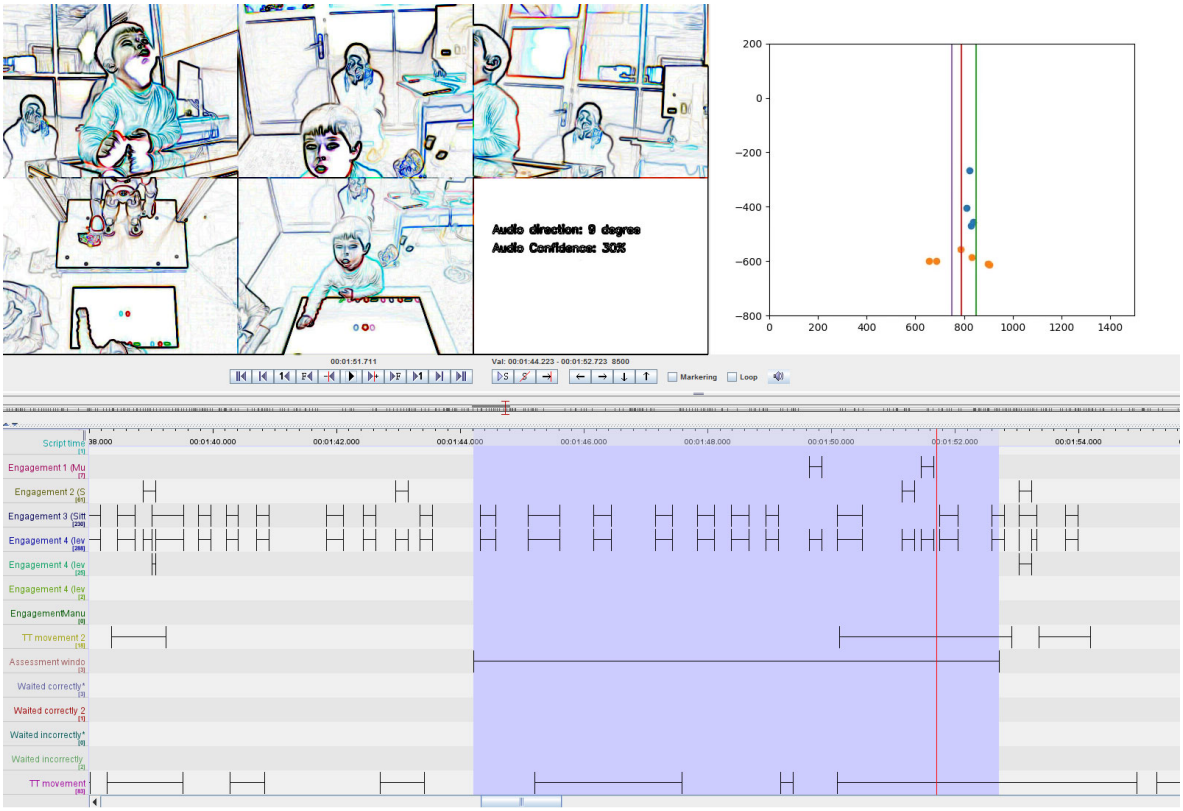

Figure 4: A screen shoot from the ELAN annotation tool. The child is moving his arm over the sand tray during the robot's turn. The duration of the robot's move is indicated in the lower part of the figure, highlighted in blue. The present time is marked by a red line.

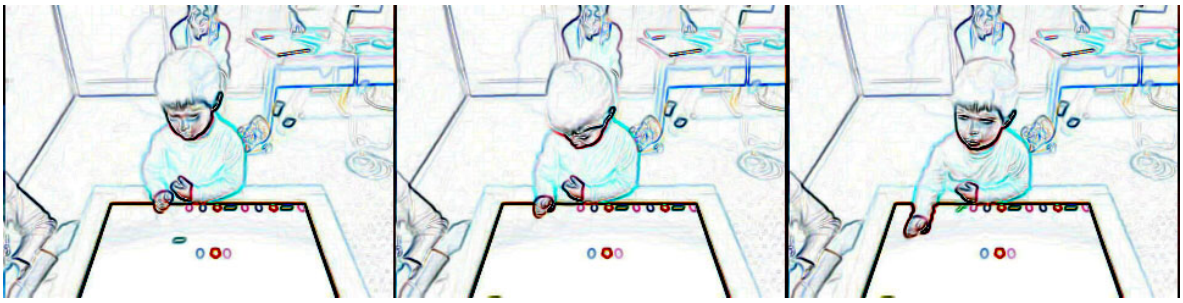

Figure 5: Three video frames from the second prior to the scene presented in Figure 4. The first, leftmost, frame displays when looking at the green object being moved by the robot. The child is waiting with his hands close to the edge of the sand tray. The middle frame displays the child looking at the object at the time it reaches it's target location. About 100 milliseconds later, displayed in the rightmost frame, the child starts to move over the sand tray.

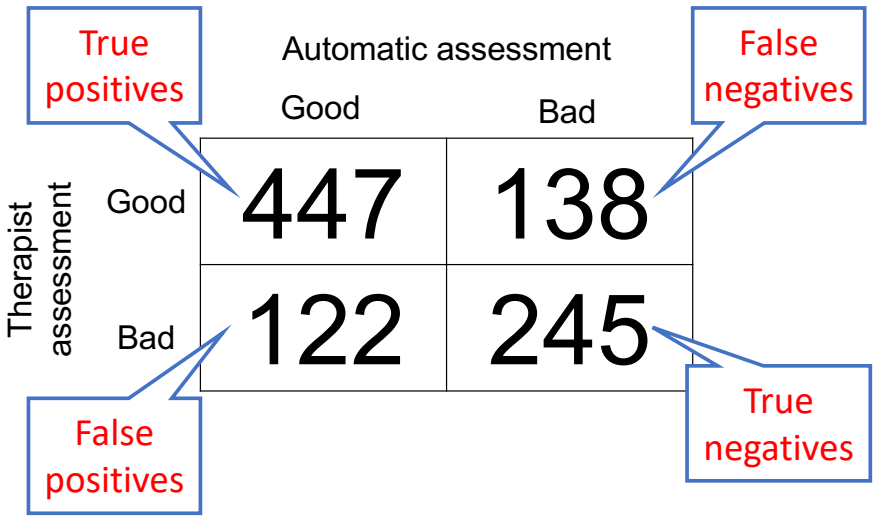

Figure 6: System assessment performance on turn-taking, using the updated methods for assessment. The agreement between human and system ratings is 73%.

offline analysis.

The algorithm has been adjusted in several ways. Firstly, only the hands and wrists are now considered during analysis, allowing the child's head and body to move over the sand tray without being interpreted as bad waiting.

Also, quick movements over the sand tray, less than 0.8 seconds, were filtered out. While this temporal filter was introduced to remove the effect of noise in the calculation of the child's hands, a closer analysis revealed that a temporal threshold effectively removed *unintentional* movements over during the opponent's turn.

Finally, as discussed in the previous section, the children sometimes touched the figures on the sand tray in the beginning of the opponent's turn, then realizes their mistake and withdraws. From a therapeutic point of view, it did not make sense to always provide negative feedback to the child in these situations and therefore classified the waiting performance as good. The effect of these adjusted assessments could be reduced by cutting the first two seconds of the assessment window, allowing the child to moving over the sand tray dung this time without receiving negative feedback. However, the more qualitative aspect of these assessments were difficult to capture and still constitute important parts of the incorrect system judgments.

The performance of the updated assessment methods is presented in 6, with an average performance of 73%.

4 Developing a Method for Exploring the Internal State Information Available in Observable Movements

This work was conducted in collaboration with Dr. Séverin Lemaignan (Bristol Robotics Lab) and Dr C E R Edmunds (University of Plymouth) who provided the dataset and contributed to the design of the experimental protocol. It is currently submitted for publication in the journal *Frontiers in Robotics & AI*.

4.1 Objectives

The first experiment was designed to establish a method for exploring what internal state information is available within a data source. Given that artificial systems can only receive an "impoverished" view of a scene (compared to human vision) it is important to establish whether the information of interest is available in the input provided to the system. Additionally, examining what information is available in a data source provides a baseline of expectation for classification systems. A large amount of research suggests that humans can use observed movements to infer the internal states of others (Gallese and Goldman, 1998; Samson, 2009; Becchio et al., 2017). An additional advantage of using this kind of information is that it can be made easily and readily available to artificial systems. As such, this experiment focuses on exploring the kinds of information that humans are able to recognize from observable movement, and thus what information we can reasonably expect an artificial system to recognize.

A lot of research has examined this question using point-light-displays and other methods to isolate movement information from other sources. This research has shown that humans are able to use movement information to infer things like gender (Hufschmidt et al., 2015), intention (Manera et al., 2010) and emotional state (Alaerts et al., 2011). Despite this, the majority of these studies use artificial stimuli. That is, the stimuli have either been constructed by filming actors instructed to behave in a certain way; e.g. with particular intentions in mind, or evoking certain emotions (Iacoboni et al., 2005; Alaerts et al., 2011). The lack of ecological validity presents a problem for HRI research. The goal of HRI is to develop systems for interacting with humans in real-world scenarios, so they need to be able to deal with natural, messy human behaviours. Consequently, a secondary goal of this experiment was to examine what kinds of information humans can recognize from naturalistic stimuli.

This study focuses on the availability of internal-state information in observable human movements and aims to evaluate the following hypothesis:

1. Humans will be able to recognize similar internal states from human movement alone as they do from the full visual scene.

To address this hypothesis participants view video clips including either the full visual scene of an interaction between two children, or clips depicting stick figures and containing only the body-pose and movement information from the same scenes. Following each clip, participants will be asked a series of questions which examine whether participants recognized any internal states (e.g. emotions, engagement) or social constructs (e.g. cooperation, dominance) based on the children's behaviours. Responses in each condition will be compared to identify which constructs are recognizable in the movement information. We expect to find that participants viewing the full scene will show higher levels of agreement in their responses, and will recognize more constructs, compared to participants viewing the movement-alone videos. However, we also expect that responses in both conditions will provide enough information to differentiate between different interactions.

4.2 Method

This study examined the effect of video type (movement-alone vs. full-scene) on responses to questions about the nature of the interaction depicted in the videos. The dataset used in this study was the PInSoRo (Lemaignan et al., 2017) dataset made openly available by our group[2]. This dataset consists of videos of child-child or child-robot interactions. Children were asked to play for as long as they wanted on a touch-screen tabletop device (henceforth: sandtray). For this study we chose to use the child-child interactions as these were more likely to involve natural social behaviours throughout the children's interactions with one another. It is important to note that interactions in this dataset were minimally controlled - pairs of children from the same school class were asked to play on a touch-screen sandtray for as long as they wanted (up to 45 minutes). Whilst structured play options were provided, they were not enforced. The stimuli used for this experiment were twenty 30-second clips. Clip selection was made based on whether they contained behaviours that could be described as either:

1. Boredom - at least one child was bored with the task on the touch-screen
2. Excitement - at least one child behaved excitedly (animated, happy)
3. Aggression - at least one child exhibited a physical aggressive action either towards the touch-screen or the other child (hitting the screen, pushing the other child's hand away)
4. Cooperation - the children were working together and/or communicating about how to perform a task
5. Dominance - one child was bossy, performing most of the actions on the touch-screen or clearly in charge
6. Aimless play - at least one child was interacting with the touch-screen in a non-goal-directed manner or without being very engaged in their task
7. Fun - at least one child was having fun (laughing, smiling)

These labels were selected based on two considerations: (a) the events and internal states (recognizable for humans) available in the dataset, and (b) events and internal states which would be useful to a robot which might observe or mediate such an interaction. Recognizing boredom and aimless behavior would allow a robot to appropriately encourage a child to take part in a task.

The movement-alone video condition was constructed by processing each clip using OpenPose[3] (Cao et al., 2017). This resulted in videos with the joint-points and lines connecting these joint-points against a black background so that each child was depicted as a stick-man-style figure.

The open and closed questions were constructed by the experimenters based on the types of social information we might want an artificial system to recognized within a scene. The open question was a single item which asked participants *"What did you notice about the interaction?"*. This question was presented immediately after the video and was designed to encourage participants to reflect on the video they had just watched. The closed questions were a series of thirty-two likert-style questions regarding either the group dynamics (e.g. *"were the children competing with one another?"*) or the behaviour of each child individually (e.g. *"was the child on the left happy?"*).

[2] https://freeplay-sandbox.github.io
[3] https://github.com/CMU-Perceptual-Computing-Lab/openpose

The experiment was designed using the jsPsych library[4], and remotely hosted from a private server (Fig. **??** shows a screenshot of the experiment). The experiment was accessible via Amazon Mechanical Turk (MTurk) to MTurk Workers. An advert was posted on MTurk containing a link to the experiment.

4.3 Analysis

4.3.1 Inter-rater Agreement

To determine inter-rater agreement and reliability, we calculated agreement scores across all 30 questions for each clip in each condition separately. This analysis was performed to examine whether participants in each condition gave similar ratings across all questions when they had viewed the same clip. High agreement would indicate that participants had interpreted similar things from a given clip, e.g. participants might all have felt that the children in a clip were being friendly and cooperative, or aggressive and competitive. Whilst this analysis does not reveal exactly what participants interpreted from the videos, it does indicate whether they gave similar ratings, and therefore reported recognizing similar states/behaviors. Given that each clip was rated by a varying subset of participants, Krippendorff's alpha (Hayes and Krippendorff, 2007) was the most appropriate metric of rater agreement (see Table 1 for number of raters and agreement per clip). The alpha scores ranged from 0.058-0.463 i.e. from "slight" to "moderate" agreement (Landis and Koch, 1977).

A t-test was conducted to assess whether the two conditions differed in their agreement scores across all 20 clips. This analysis revealed that participants in the full-scene condition showed significantly higher agreement ($M = 0.328$, $SD = 0.110$) than participants in the movement-alone condition ($M = 0.252$, $SD = 0.079$) (Paired Samples T-Test: $t(197) = 2.95$, $p = 0.008$, $d = 0.78$). These analyses show that participants viewing the full-scene clips demonstrated higher levels of agreement in their ratings than those viewing the movement-alone clips. However, participants in the latter condition still showed some agreement (chance level Krippendorff's Alpha $= 0.0$), suggesting that some social constructs were recognizable within the movement information in both conditions.

We then trained a classifier on the full-scene data with hand-crafted social labels to then attempt to automatically identify these social labels on the movement-alone data. Our best performing classifier (a 3-kNN) on 80% of the full-scene data and testing on the remaining 20% results in a (cross-validated) precision of 46.2% and recall of 33.6%. We found very similar levels of precision and recall (respectively 41.6% and 32.7%) when testing on the movement-alone ratings: the assessment of the social interaction taking place between two children, made by naive observers watching a low-dimensional, movement-alone video-clip of the interaction, carries similar informational content regarding the social interaction as the original raw video footage.

Based on this finding we can tentatively conclude that the movement-alone data contains a comparable amount of information as the full-scene data, with respect to social interactions. Furthermore, this information, contained in the movements of interacting individuals, can be interpreted by human observers in a similar way as when the observers are viewing the full visual scene.

While above chance, the accuracy of the classifier is relatively low. This may reflect the inherent difficulty of rating social experiences for an external, naive observer (such as the raters recruited for this study).

To better make sense of these results, we employed a second data mining technique (Exploratory Factor Analysis, EFA) to attempt to uncover underlying latent factors that would, in effect, embody

[4] https://www.jspsych.org/

Table 1: Table of inter-rater agreement scores for responses to each clip in each condition

Clip	Krippendorff's Alpha (3 d.p.)	
	Full-Scene (N)	Movement Alone (N)
1	0.446 (16)	0.186 (26)
2	0.181 (24)	0.270 (20)
3	0.393 (22)	0.369 (18)
4	0.444 (22)	0.262 (23)
5	0.328 (23)	0.283 (20)
6	0.463 (19)	0.359 (19)
7	0.091 (19)	0.236 (23)
8	0.339 (19)	0.312 (17)
9	0.097 (20)	0.058 (18)
10	0.396 (18)	0.086 (13)
11	0.280 (17)	0.234 (23)
12	0.368 (25)	0.298 (16)
13	0.334 (20)	0.189 (21)
14	0.310 (17)	0.309 (21)
15	0.422 (26)	0.242 (14)
16	0.192 (16)	0.272 (21)
17	0.273 (17)	0.183 (21)
18	0.334 (16)	0.331 (24)
19	0.415 (22)	0.304 (19)
20	0.451 (18)	0.250 (23)

stronger cognitive constructs, implicitly relied upon by the humans when assessing a social interaction. We ran independent EFAs on the ratings provided for the full-scene videos and those provided for the movement-alone clips. In both condition, one factor was measuring the **behavioral imbalance** between the two children (how similar or dissimilar their behaviors were); a second factor reflected the **valence of the interaction**, from adversarial behaviors and negative emotions, to pro-social and positive behaviors and emotions; finally a third factor embodied **the level of engagement** of the children. These constructs may be indicative of the constructs humans use to interpret social situations in general.

The results of both the classification analysis and EFA demonstrate that it is reasonable to expect a machine learning algorithm, and in consequence, a robot, to successfully decode and classify a range of social situations using a low-dimensional data source (such as the movements and poses of observed individuals) as input: our study shows that, even thought assessing social interactions is difficult even for humans, using skeletons and facial landmarks only does not significantly degrades the assessment.

5 Proof-of-Concept of a Conceptor-Based System for Classifying Internal States from Observable Movements

5.1 Objectives

Next, we wish to evaluate the usefulness of conceptors for learning and classifying human internal states based on observable behaviour. This study is a proof-of-concept study for the conceptor-based system which is intended to be applied to the problem of classifying the behaviours of children with ASD in terms of different "levels" of severity. This study will also form the basis for a series of studies aiming to develop a method for representing and classifying human internal states in terms of continuous dimensions rather than discrete categories. This has been published as a workshop paper at HRI 2019.

Conceptors are neurocomputational mechanisms capable of learning dynamical patterns (Jaeger, 2017). The fact that conceptors can be used to learn a large number of patterns based on only the prototypical extremes of a continuum encompasses the two main features which we wish to exploit for this project. That is: (1) we want to be able to train a system to recognize a large number of patterns using a minimal training dataset, and (2) we want the combination of these patterns/categories to reflect a continuous dimension. As such, the main aims of this proof-of-concept study are two-fold. Firstly we explore whether conceptors can be trained to recognize the extremes of a human internal state from observable behaviour. Secondly, once these extremes are learned we aim to generate a conceptor representing an intermediate level of this internal state by combining the learned conceptors. We will then test whether this new conceptor can be used to recognize the intermediate state without being trained.

For this study we have operationalized "internal state" as a state which a human experiences but does not intentionally communicate. As such, we needed a dataset depicting individuals in isolation who were experiencing varying intensities of an internal state. In order to avoid emotional states, which are being extensively examined by similar studies and are usually highly communicative, we chose to focus on the internal state of task engagement. We therefore aim to evaluate the following hypotheses:

1. A conceptor-based system can be trained to recognize high and low engagement based on human movement data

2. Once the extremes are learned, the resultant conceptors can be combined to create a representation of an intermediate state of task engagement

3. The intermediate state conceptor can be used to classify human movement data without needing to be trained

5.2 Dataset

The dataset for this study consisted of clips taken from the PInSoRo dataset (Lemaignan et al., 2017)[5]. This dataset was chosen for a number of reasons, one being that many of the videos have already been annotated with labels approximating task engagement. Below are the definitions of the labels we used (see Lemaignan et al. (2018)):

- **Goal-Oriented Play** - defined as purposeful, structured play involving some action planning. We argue that this reflects high task engagement.
- **Aimless Play** - defined as unstructured or silly play. We argue that this reflects and intermediate level of task engagement.
- **No Play** - defined as a lack of any obvious activity. We argue that this reflects low task engagement.

We selected clips of child-robot interactions as we believe that the children were less likely to be actively/intentionally communicating their engagement state to the robot.

We extracted all the clips which had been annotated with these labels. The clips were then processed using OpenPose (Cao et al., 2017) to obtain a file containing the xyz coordinates for the child's joints and facial features for each frame. These sequences of coordinates were used as input for the conceptor-based classifier. The training dataset was made up of a subset of the "high" and "low" engagement clips. The remaining clips with these labels were used as a test dataset. Additionally, all of the examples of "intermediate" engagement were kept as test stimuli to see whether a new conceptor, generated by combining the trained conceptors, could be used to classify these stimuli without having been trained.

5.3 Validation of the labels

We also briefly validated the labels to ensure their utility. To this end, five participants (students and employees) were recruited from the University of Plymouth's School of Computing, Electronics and Mathematics on a volunteer basis.

A total of forty-five video clips were extracted from the data set for this study. We selected fifteen clips with the annotation "goal-oriented play", fifteen with the annotation "aimless play" and fifteen with the annotation "no play". Clip lengths ranged from 12-30 seconds. After clips were selected we extracted both the full visual scene versions and the movement-alone versions. The movement-alone versions were processed such that they depicted the children's joint-points, connected by coloured lines, against a black background. These videos act as visual representations of the data used as input for the conceptor-based system in that they depict only movement and pose information by showing the position of the child's body in each frame.

[5] https://freeplay-sandbox.github.io

Participants watched the full visual scene videos on one day and were then asked to return the next day when they would watch the movement-alone videos. Participants all received the following instructions before beginning the experiment:

You're about to watch several videos of children interacting with a touch-screen table-top. The children were able to either play a specific game on the touch-screen, or to do whatever they want. After each clip you will be asked to judge the child's level of task engagement.

Participants then viewed nine of each type of clip (a total of twenty-seven clips) presented in a random order. Following each clip, participants were presented with the question "*How engaged was the child with their task on the touch screen table-top?*". This question was accompanied by a 7-point Likert scale ranging from 1 = "Not at all Engaged" to 7 = "Highly Engaged". Participants used this scale to report how engaged they thought the child in the clip had been and then continued on to the next clip. On the second day, the experiment proceeded in the same way except participants were shown the movement-alone videos instead of the full visual scene videos. Each participant saw the same twenty-seven clips in both sessions.

5.3.1 Inter-rater agreement

We firstly examined inter-rater agreement by calculating Krippendorff's alpha for the responses. We initially checked whether participants gave similar responses for each of the three types of videos. To do this, Krippendorff's alpha was calculated for responses to all of the videos of each type. The alpha scores have been interpreted in terms of the benchmarks outlined by Landis and Koch **?**. Responses showed "fair" agreement for the goal-oriented (high engagement) clips (Krippendorff's alpha = 0.269) and the no-play (low engagement) clips (Krippendorff's alpha = 0.267). Responses for aimless (intermediate engagement) clips showed "slight" agreement (Krippendorff's alpha = 0.171). The low levels of agreement can partially be explained by the fact that there were very few raters (2-4) per clip. As such we did not expect perfect levels of agreement and argue that the levels obtained suggest a sufficient degree of similarity in participants' ratings.

We then examined whether participants had higher agreement when viewing the full visual scene clips compared to the movement-alone clips for each clip type. The results of this analysis are reported in Table 2. For the goal-oriented and no-play clips, participants tended to show similar levels of agreement in each condition. However, for the aimless clips, participants demonstrated poor agreement when viewing the movement-alone clips.

Table 2: Table of inter-rater agreement scores for responses to each clip-type in each condition

Clip Type	Krippendorff's Alpha (3 d.p.)	
	Full Scene	movement-alone
Goal Oriented	0.382 (fair)	0.368 (fair)
Aimless	0.247 (fair)	-0.022 (poor)
No Play	0.126 (slight)	0.202 (fair)

5.3.2 Ratings

The second set of analyses looked at the how participants rated each type of video. Overall mean rating was 4.81 (SD = 1.25) for goal-oriented clips, 4.16 (SD = 1.52) for aimless clips, and 2.43 (SD = 1.54) for no-play clips. An ANOVA revealed a significant main effect of clip-type on ratings ($F(2,267)=64.99$, $p<0.001$). Importantly, a *post hoc* Tukey test revealed significant differences between all conditions (Tukeys HSD: all differences >0.6, all ps <0.007; see Table 3).

Table 3: Table of results for post hoc Tukey's Honest Significant Difference test.

Comparison	Difference	Significance (p adj)
Goal Oriented – Aimless	0.656	$p = 0.007$
Goal Oriented – No Play	2.348	$p < 0.001$
Aimless – No Play	1.722	$p < 0.001$

5.3.3 Label validity

These results demonstrate that participants rated the clips in terms of engagement such that goal-oriented clips showed the highest levels of engagement, no-play clips showed the lowest levels, and aimless clips fell in-between these two extremes. Consequently, we feel our assumption that these annotations reflect different levels of engagement is sufficiently supported for these data to be used to train and test a conceptor-based classifier to recognize engagement based on observable behaviour. The remainder of this paper describes the design and initial tests of such a classifier.

5.4 Conceptor-Based Classifier

The conceptor-based approach is based on the phenomenon in Recurrent Neural Networks whereby when a reservoir is driven by a pattern, the resultant network states are confined to a linear subspace of the network state space which is characteristic of that pattern (Jaeger, 2014). As a result, for this study, the overt behaviours which characterize the different levels (classes) of engagement will occupy different regions of the state space, and can be encoded in a conceptor. This conceptor (C_j) then acts as a map associated with a particular pattern (p_j).

Building this conceptor-based classifier involves computing J conceptors, one for each class of engagement. Obtaining these conceptors can be broken down into several steps. First, an Echo State Network (ESN) is created with an input layer of K input units and a hidden layer reservoir of N neurons. For each class the network was driven, independently, with all training samples s_j^m in each class j, according to the ESN state update equation:

$$x(n+1) = \tanh(W \cdot x(n) + W^{in} \cdot p(n+1) + b) \tag{1}$$

This yielded a set of network states $X_j = [x(1) \dots x(t)]$ where t was the number of time-steps in s_j from which a state correlation matrix $R_j = X_j X_j^T / M_j$ could be obtained, where M_j is the total number of samples for class j. Next we compute conceptor C_j through the equation:

$$C(R, \alpha) = R(R + \alpha^{-2})^{-1} \tag{2}$$

Table 4: Predicting internal states with Conceptors.

Algorithm: Conceptor-based classification.
Input: A test sample s belonging to one a class j.

1. Take a sample s from the test set.
2. Drive the reservoir with sample s to obtain a state vector $z = [x(1) \cdots x(n)]$, where n is the $\#$ of steps in s.
3. For each Conceptor C_j compute $h(j) = z^T C_j z$, a "positive evidence" quantity of z belonging to class j.
4. Collect each evidence $h(j)$ into a $j-$dimensional classification hypothesis vector $h^+ = \{h(1) \cdots h(j)\}$.
5. Classify s as belonging to class j from $j = argmax(h^+)$.
6. **END**

Output: Class sample s belongs to.

Where R is a correlation matrix and $\alpha \in (0, \infty)$ in an "apperture" parameter. For more see Jaeger (2014).

Once a conceptor matrix was computed for each class, a new sample s from the test set could be classified by feeding it into the ESN reservoir and obtaining a new state vector $z = [x(1) \dots x(n)]$. Then, for each conceptor a "positive evidence quantity $z^T C_j z$ was computed. Classification is then done by deciding for $j = argmax(z^T C_j z)$ where the class j is the one to which the sample s belongs.

The procedure for the conceptor-based classifier is summarize in Table 4.

5.5 Results

The resultant conceptors were tested using previously unseen samples from the high and low engagement categories. The results of this test are shown in Figure 7. Performance is above chance for both classes (high engagement: 60%, low engagement: 75%).

The conceptor-based system thus successfully learned to recognize high and low engagement from observable human movement. Future work will construct new conceptors by linearly combining these learned conceptors. We will then test whether these new conceptors can be used to recognize intermediate levels of engagement identified in the PInSoRo data set.

If new conceptors can be generated, this method will show promise for use in providing diagnostic information for clinicians assessing children with ASD. The ability to interpolate between extremes along a continuum means that such a system could be trained on a smaller dataset, whilst still achieving a high level of detail through the generation of multiple intermediate classification categories.

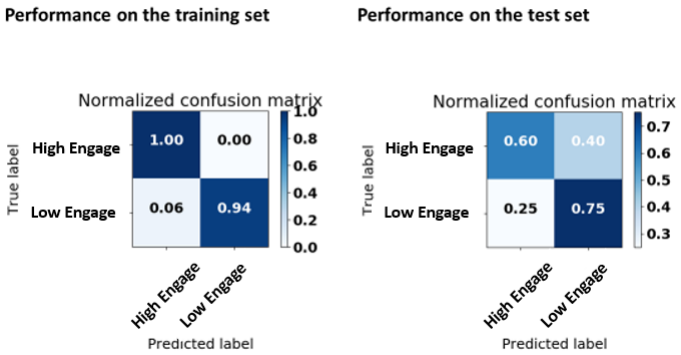

Figure 7: Confusion matrices showing classification performance of trained conceptors on training data (left) and test data (right).

6 Work to be continued beyond the conclusion of DREAM

6.1 Continue to develop the conceptor based system for ASD diagnosis

Although the proof of concept has so far been encouraging, we still need to apply the conceptor-based system to the problem of providing a description of a case of Autism Spectrum Disorder (ASD) in terms of severity via an automated behaviour recognition system. Depending on a number of factors this experiment could adopt one of two approaches. The first, and most favourable, approach will be to use an adapted version of the protocol described before to explore what features of a child's behaviour clinicians most commonly use to identify severity. This will make use of the video data from the DREAM project (DREAM, 2014).

Clinicians will be presented with short clips of children taking part in intervention settings and asked to describe the child's behaviour and indicate any diagnostic thoughts or conclusions they may have drawn from each clip. Whilst the final diagnosis of each child is also provided by the DREAM project for this data set, we would use this protocol to explore the symptomology of each case individually. This approach incorporates the current understanding of ASD as a highly heterogeneous diagnosis whereby the severity and presentation of symptoms differs from case to case. Consequently, creating an automated diagnostic system able to provide information about the severity of different symptoms requires a highly detailed training data set. Additionally, we would need to identify which, if any, symptoms we are able to identify based on observable behaviours. This study would achieve this by asking clinicians to be specific about how they feel the child's behaviours may relate to any diagnostic conclusions they made based on the video data. Once a set of behaviours has been identified, as well as how they relate to symptomology or severity, a dataset will be constructed to train the conceptor-based system on high and low severity behaviours. Conceptors for classifying intermediate severity levels will be generated using linear combinations of the extremes of each behaviour. The system will then be tested on the remaining data set to assess its accuracy compared to the severity scores and symptom descriptions offered by the clinicians. Due to how extensive this study would need to be, and the potential difficulty of recruited a large number of highly specialized participants, we appreciate that this study may not be feasible within the given time-frame. If this becomes the case, a secondary approach will be used.

The alternative approach will be to simply train the conceptor-based system on examples of behaviour from children who have received a final diagnosed of either high or low severity ASD. This experiment will also use the video and diagnostic data available in the DREAM dataset. This process

will allow us to explore what behavioural features the conceptor-based system uses to differentiate between each severity class. Similarly to the first approach, once conceptors have been trained for the extreme classes, new, intermediate conceptors will be generated for intermediate levels of severity. The system will then be tested on unseen examples of the extremes, as well as examples of children with intermediate severity diagnoses. Judgements of the system's accuracy will be made by comparing the system's severity classification to the final diagnosis in the DREAM dataset.

The aim for both of these approaches will be to develop a conceptor-based system able to provide diagnostic information about children with ASD. Specifically, we want to create a system which meets the following requirements:

1. Can provide multiple classification options ranging from "low severity" to "high severity", avoiding a reductionist view of ASD.

2. Achieves 1. whilst requiring a relatively small training data set.

6.2 Conceptor alternative using the neuro-engineering framework

Given that the performance of the conceptors have room for improvement, we will also explore an alternative approach to learning numerous classes of dynamical pattern based on the prototypical extremes. Given that the approach of representing human internal states as continuous dimensions is inspired by definitions of these states provided by Psychology, we consider an approach which has the potential to more closely reflect how humans represent, and even recognize, human internal states. For this experiment we will develop an equivalent to the conceptor-based system using the biologically plausible neural simulator Nengo (Bekolay et al., 2014).

Recently, Voelker and Eliasmith (2018) have presented a biologically plausible dynamical spiking neural network, formulated in terms of the so-called Neural Engineering Framework (Eliasmith and Anderson, 2003) – which is the framework underlying Nengo – capable of exactly reproducing delayed time signals. Their approach can be understood as solving the question of what an optimal reservoir would be for a given problem, achieved by starting with a mathematical description of the ideal system, from which the network as such is then derived. In other words, it achieves an optimal solution for the problem that Echo State Networks, and consequently concpetors, approximate. However, this network has so far only been demonstrated on very abstract, single-dimensional input patterns, whereas conceptors easily encode multi-dimensional imputs (such as the locations of various point-light markers on a human skeleton), as we have demonstrated above.

The approach of Voelker and Eliasmith (2018) is promising as a possible alternative to conceptors since the mathematical formulation leads us to expect higher levels of accuracy and performance. Additionally, given that many of the present-day and near-future applications lie in human-machine interaction, the spiking nature of these networks (compared with the rate-based approach in Echo State Networks and Conceptors), in combination with neuromorphic hardware, is potentially interesting since real-time performance can thus be guaranteed (assuming sufficient hardware) while there are also results indicating that such a network would be more energy-efficient Blouw et al. (2018). It has, however, not been explored how such a network would encode signals of dimensionality higher than one, and it is not clear that inputs that would be interesting in real-life conditions can be reduced to a single dimension.

We have obtained initial results that suggest reasonable classification performance using just two of the principal components obtained from a PCA on the PinSoRo data-set, and can additionally demonstrate that a system trained to recognise high or low engagement as extrema will classify exam-

ples of intermediate pattern as also falling between these extrema. The work is therefore promising, but needs to be continued before a final decision can be made.

This Nengo-based approach will then be compared directly with the conceptor-based approach to assess similarities, differences and whether one system improves upon the other. The main aim for this experiment will be to explore an alternative way of addressing the system requirements outlined above.

References

Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., and Wenderoth, N. (2011). Action and emotion recognition from point light displays: an investigation of gender differences. *PloS one*, 6(6):e20989.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition.

Becchio, C., Koul, A., Ansuini, C., Bertone, C., and Cavallo, A. (2017). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of life reviews*.

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., Choo, X., Voelker, A., and Eliasmith, C. (2014). Nengo: a python tool for building large-scale functional brain models. *Frontiers in neuroinformatics*, 7:48.

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E. J., Kopp, S., Bergmann, K., Leseman, P., Küntay, A. C., Göksun, T., et al. (2015). L2TOR-second language tutoring using social robots. In *Proceedings of the ICSR 2015 WONDER Workshop*.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.

Blouw, P., Choo, X., Hunsberger, E., and Eliasmith, C. (2018). Benchmarking keyword spotting efficiency on neuromorphic hardware. *CoRR*, abs/1812.01739.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., and Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of Autism and Developmental Disorders*, 45(5):1121–1136.

Cai, H., Fang, Y., Ju, Z., Costescu, C., David, D., Billing, E. A., Ziemke, T., Thill, S., Belpaeme, T., Vanderborght, B., Vernon, D., Richardson, K., and Liu, H. (2018). Sensing-enhanced Therapy System for Assessing Children with Autism Spectrum Disorders: A Feasibility Study. *IEEE Sensors Journal*.

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

Clarke, T. J., Bradshaw, M. F., Field, D. T., Hampson, S. E., and Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34(10):1171–1180.

Cohen, P. A., Kulik, J. A., and Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American educational research journal*, 19(2):237–248.

Daoudi, M., Coello, Y., Desrosiers, P., and Ott, L. (2018 May). A New Computational Approach to Identify Human Social intention in Action. *In IEEE International Conference on Automatic Face and Gesture Recognition*.

De Silva, R. S., Tadano, K., Higashi, M., Saito, A., and Lambacher, S. G. (2009). Therapeutic-assisted robot for children with autism. *In Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, Oct:3561–3567.

Department for Education (2018). Schools, pupils and their characteristics: January 2018. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/719226/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf`, Last accessed on 21-11-2018.

DREAM (2014). Goals methodology — DREAM project. `https://www.dream2020.eu/goals-methodology/`.

Duquette, A., Michaud, F., and Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low- functioning autism. *Autonomous Robots*, 24(2):147–157.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Eliasmith, C. and Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems.* MIT Press, Cambridge, MA.

Esteban, P. G., Baxter, P., Belpaeme, T., Billing, E., Cai, H., Cao, H. L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., and Fang, Y. (2017). How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn, Journal of Behavioral Robotics*, 8(1):18–38.

Eysenck, H. J. (1950). *Dimensions of personality.*, volume 5. Transaction Publishers.

Franois, D., Polani, D., and Dautenhahn, K. (2007). On-line behaviour classification and adaptation to human-robot interaction styles. *In Proceedings of the ACM/IEEE international conference on Human-robot interaction*, March:295–302.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2):593–609.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12):493–501.

Geschwind, D. H. and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology*, 17(1):103–111.

Goldman, A. I. et al. (2012). Theory of mind. *The Oxford handbook of philosophy of cognitive science*, pages 402–424.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). The cowriter project: Teaching a robot how to write. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 269–269. ACM.

Hufschmidt, C., Weege, B., Rder, S., Pisanski, K., Neave, N., and Fink, B. (2015). Physical strength and gender identification from dance movements. *Personality and Individual Differences*, 76:13–17.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS biology*, 3(3):e79.

Jaeger, H. (2014). Conceptors: an easy introduction. *arXiv preprint arXiv:1406.2671*.

Jaeger, H. (2017). Using conceptors to manage neural long-term memories for temporal patterns. *Journal of Machine Learning Research*, 18(13):1–43.

Kanade, T., Tian, Y., and Cohn, J. F. (2000). Comprehensive database for facial expression analysis. In *fg*, page 46. IEEE.

Kelley, R., Tavakkoli, A., King, C., Nicolescu, M., Nicolescu, M., and Bebis, G. (2008). Understanding human intentions via hidden markov models in autonomous mobile robots. *Proceedings of the 3rd international conference on Human robot interaction - HRI '08*, pages 367–374.

Lai, M. C., Lombardo, M. V., Hakrabarti, B., and Baron-Cohen, S. (2013). Subgrouping the Autism Spectrum: Reflections on DSM-5. *PLoS Biology*, 11(4):e1001544.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lemaignan, S., Edmunds, C. E. R., Senft, E., and Belpaeme, T. (2017). The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a Dataset of Social Interactions. *arXiv preprint arXiv:1712.02421*.

Lemaignan, S., Edmunds, C. E. R., Senft, E., and Belpaeme, T. (2018). The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one*, 13(10):e0205999.

Lewkowicz, D., Delevoye-Turrell, Y., Bailly, D., Andry, P., and Gaussier, P. (2013). Reading motor intention through mental imagery. *Adaptive Behavior*, 21(5):315–327.

Leyzberg, D., Spaulding, S., and Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 423–430. ACM.

Lord, C., Risi, S., Lambrecht, L., Cook Jr, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation ScheduleGeneric: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, 30(3).

Manera, V., Becchio, C., Cavallo, A., Sartori, L., and Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research*, 211(3-4):547–556.

Manera, V., Schouten, B., Becchio, C., Bara, B. G., and Verfaillie, K. (2010). Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behavior research methods*, 42(1):168–178.

Ozonoff, S. (1995). Reliability and validity of theWisconsin Card Sorting Test in studies of autism. *Neuropsychology*, 9(4):491.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Ricks, D. J. and Colton, M. B. (2010). Trends and considerations in robot-assisted autism therapy. *In Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 2010:4354–4359.

Robins, B., Dautenhahn, K., Te Boekhorst, R., and Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120.

Rogers, C. L., Goddard, L., Hill, E. L., Henry, L. A., and Crane, L. (2016). Experiences of diagnosing autism spectrum disorder: a survey of professionals in the United Kingdom. *Autism*, 20(7):820–831.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Samson, D. (2009). Reading other people's mind: Insights from neuropsychology. *Journal of Neuropsychology*, 3(1):3–16.

Scassellati, B. (2005a). Using social robots to study abnormal social development. *5th International Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotic Systems, Lund University Cognitive Studies*, pages 11–14.

Scassellati, B. (2005b). Quantitative metrics of social response for autism diagnosis. *In Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop.*, 2005(September 2005):585–590.

Scassellati, B., Admoni, H., and Matarić, M. (2012). Robots for Use in Autism Research. *Annual Review of Biomedical Engineering*, 14(1):275–294.

Thill, S., Pop, C. A., Belpaeme, T., Ziemke, T., and Vanderborght, B. (2012). Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn*, 3(4):209–217.

Voelker, A. R. and Eliasmith, C. (2018). Improving spiking dynamical networks: Accurate delays, higher-order synapses, and time cells. *Neural Computation*, 30(3):569–609.

Wall, D. P., Dally, R., Luyster, R., Jung, J. Y., and DeLuca, T. F. (2012b). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS one*, 7(8):e43855.

Wall, D. P., Kosmicki, J., Deluca, T. F., Harstad, E., and Fusaro, V. A. (2012a). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry*, 2(4):e100.

Wimmer, M., MacDonald, B. A., d. Jayamuni, and Yadav, A. (2008). Facial expression recognition for human-robot interaction–a prototype. In *International Workshop on Robot Vision*, pages 139–152. Springer.

Zunino, A., Morerio, P., Cavallo, A., Ansuini, C., Podda, J., Battaglia, F., Veneselli, E., Becchio, C., and Murino, V. (2018). Video Gesture Analysis for Autism Spectrum Disorder Detection. https://www.researchgate.net/profile/Andrea_Zunino2/publication/327751352_Video_Gesture_Analysis_for_Autism_Spectrum_Disorder_Detection/links/5ba25bd792851ca9ed15b1c9/Video-Gesture-Analysis-for-Autism-Spectrum-Disorder-Detection.pdf, Last accessed on 19-11-2018.