Research Article **Open Access**

Pablo G. Esteban*, Paul Baxter, Tony Belpaeme, Erik Billing, Haibin Cai, Hoang-Long Cao, Mark Coeckelbergh, Cristina Costescu, Daniel David, Albert De Beir, Yinfeng Fang, Zhaojie Ju, James Kennedy, Honghai Liu, Alexandre Mazel, Amit Pandey, Kathleen Richardson, Emmanuel Senft, Serge Thill, Greet Van de Perre, Bram Vanderborght, David Vernon, Hui Yu, and Tom Ziemke

How to Build a Supervised Autonomous System for Robot-Enhanced Therapy for Children with Autism Spectrum Disorder

DOI 10.1515/pjbr-2017-0002
Received August 5, 2016; accepted April 9, 2017

Abstract: Robot-Assisted Therapy (RAT) has successfully been used to improve social skills in children with autism spectrum disorders (ASD) through remote control of the robot in so-called Wizard of Oz (WoZ) paradigms. However, there is a need to increase the autonomy of the robot both to lighten the burden on human therapists (who have to remain in control and, importantly, supervise the robot) and to provide a consistent therapeutic experience. This paper seeks to provide insight into increasing the autonomy level of social robots in therapy to move beyond WoZ. With the final aim of improved human-human social interaction for the children, this multidisciplinary research seeks to facilitate the use of social robots as tools in clinical situations by addressing the challenge of increasing robot autonomy. We introduce the clinical framework in which the developments are tested, alongside initial data obtained from patients in a first phase of the project using a WoZ set-up mimicking the targeted supervised-autonomy behaviour. We further describe the implemented system architecture capable of providing the robot with supervised autonomy.

Keywords: Robot-Enhanced Therapy, Autism Spectrum Disorders, Supervised Autonomy, Multi-sensory Data, Cognitive Controller

1 Introduction

Autism Spectrum Disorder (ASD) is characterised by impairments in social interactions and communication, usually accompanied by restricted interests and repetitive behaviour [1]. Most individuals with ASD require professional care throughout their lives [2, 3], entailing a significant financial and time (at least 15 hours per week) commitment [4, 5].

Evidence-based psychotherapy necessitates both clinical expertise and expertise in applying the results of scientific studies. For ASD, one of the most efficient ways of improving individuals' abilities and reducing their symptoms is through early (cognitive-) behavioural intervention programs [6]. Studies testing the effectiveness of such interventions report significant results in terms of language and social skill improvement, decreased stereotypical behaviours, and acceleration of developmental rates [7].

***Corresponding Author: Pablo G. Esteban:** Robotics and Multibody Mechanics Research Group, Agile & Human Centered Production and Robotic Systems Research Priority of Flanders Make, Vrije Universiteit Brussel, Brussels, Belgium, E-mail: pablo.gomez.esteban@vub.be
Hoang-Long Cao, Albert De Beir, Greet Van de Perre, Bram Vanderborght: Robotics and Multibody Mechanics Research Group, Agile & Human Centered Production and Robotic Systems Research Priority of Flanders Make, Vrije Universiteit Brussel, Brussels, Belgium
Erik Billing, Serge Thill, David Vernon, Tom Ziemke: Interaction Lab School of Informatics, University of Skövde, Skövde, Sweden
Cristina Costescu, Daniel David: Department of Clinical Psychology and Psychotherapy, Babeş-Bolyai University, Cluj-Napoca, Romania
Paul Baxter, Tony Belpaeme, James Kennedy, Emmanuel Senft: Centre for Robotics and Neural Systems, Plymouth University, Plymouth, U.K.
Haibin Cai, Yinfeng Fang, Zhaojie Ju, Honghai Liu, Hui Yu: School of Computing, University of Portsmouth, Portsmouth, U.K.
Mark Coeckelbergh, Kathleen Richardson: Centre for Computing and Social Responsibility, Faculty of Technology, De Montfort University, Leicester, U.K.
Alexandre Mazel, Amit Pandey: Softbank Robotics Europe, Paris, France

Although behavioural approaches have demonstrated effectiveness in reducing ASD symptoms, there is more to be done in this field. It is important to improve the efficiency of early behavioural interventions to ensure progress at a later stage, allowing adults with ASD to lead independent (or near-independent) lives [8]. Taking into account that individuals with ASD tend to be more responsive to feedback coming from an interaction with technology rather than from an interaction with human beings [9], and given the need for reducing costs while increasing the effectiveness of standard (cognitive-) behavioural therapies, studies have shown that robots may be beneficial in ASD therapies as mediators between human models and ASD children, see [9–11]. In the Robo-Mediator approach [12], a social robot is used as a means of delivering the standard treatment, elucidating faster and greater gains from the therapeutic intervention when compared to classical treatment. Several robots have already been used in Robot-Assisted Therapy (RAT) with children with ASD: the NAO robot, see [13–15] among others; low-cost robots like AiSOY1 [16] or CHARLIE [17]; robots that use their touchscreens as part of the interaction like CARO and iRobiQ [18]; or the robot Probo which has been used for social story telling [19], to improve play skills [20], and to mediate social play skills of children with ASD with their sibling (brother or sister) [21]. See [22] for a complete survey detailing how RAT robots are mapped to therapeutic and educational objectives.

1.1 Increasing autonomy in RAT

Typical work in RAT is performed using remote controlled robots; a technique called Wizard of Oz (WoZ) [23, 24]. The robot is usually controlled, unbeknownst to the child, by another human operator. This permits the therapists to focus on achieving a higher level of social interaction without requiring sophisticated systems reliant on high levels of artificial intelligence. However, WoZ is not a sustainable technique in the long term, see [25]. It is a costly procedure as it requires the robot to be operated by an additional person and as the robot is not recording the performance during the therapy, additional time resources are needed after the intervention.

It has been proposed that robots in future therapeutic scenarios should be capable of operating autonomously (while remaining under the supervision of the therapist) for at least some of the time [26]. Providing the robots with autonomy in this sense has the potential to lighten the therapist's burden, not only in the therapeutic session itself but also in longer-term diagnostic tasks. Indeed, as this paper will illustrate, the technical solutions required to deliver adequate autonomous abilities can also be used to improve diagnostic tools, for example by collecting quantitative data from the interaction, or automatically annotating videos of interactions with the children (currently a manual process involving significant time and effort by multiple therapists [25]). Diagnosis might further be improved through automated behaviour evaluation systems (required to allow the robot to choose appropriate actions during autonomous behaviour).

A system capable of such data processing can help therapists to administer personalised interventions for each child, as the robot could infer intentions, needs, or even the mood of the child based on previous interactions [26]. A fully autonomous robot might be able to infer and interpret a child's intentions in order to understand their behaviour and provide real-time adaptive behaviour given that child's individual needs. An autonomous robot could attempt to (re-)engage the child should they lose interest in the therapeutic task. Robots also need to respond to high level commands from therapists, enabling the latter to overrule the robot behaviour at any time. Such a degree of autonomy would enable the development of less structured interaction environments which may help to keep the child engaged [27], e.g., by providing the child with ASD the ability to make choices during the interaction with the robot [28]. A high level of engagement would be reinforced by explicit positive feedback as it has been proven that children find rewards particularly encouraging, see [29, 30]. This encouragement can be given in the form of sensory rewards, such as the robot clapping hands or playing some music.

In building more autonomous robots capable of flexible social behaviour, the development of a basic "intentional stance" [31] is important. In ideal circumstances, this means that the robot should be able to take a perspective on the mental state of the child with whom it is interacting, i.e., it should be able to develop a Theory of Mind (ToM) and be able to learn from normal social signals in a manner that is similar to the way humans learn to infer the mental states of others. Full autonomy (in the sense that the robot can adapt to any event during the therapeutic sessions) is currently unrealistic and not desired as the robot's action policy will not be perfect and in a therapeutic scenario, every single action executed by the robot should be appropriate to the therapeutic goals, context of the interaction, and state of the child. However it is feasible to aim for a "supervised autonomy", where the robot user (the therapist, psychologist or teacher) gives the robot particular goals and the robot autonomously

works towards achieving these goals whilst allowing the supervisor to override every action prior to execution to ensure that only therapeutically valid actions are executed.

Increasing the autonomy of robots will also bring about a new set of challenges. In particular, there will be a need to answer new ethical questions regarding the use of robots with vulnerable children, as well as a need to ensure ethically-compliant robot behaviour (e.g., to avoid persisting with certain behaviour should the child refuse to collaborate).

Architectures for controlling autonomous social robots commonly utilise behaviour-based architectures, as these systems are capable of mixing different behaviours and being responsive to external sensory information [32]. However, these approaches operate in-the-moment and are not capable of anticipating upcoming events, which might be desirable when interacting with ASD children. Few of the existing control architectures are tailored to social robotics for therapeutic purposes. B3IA [33] is a control architecture for autonomous robot-assisted behavioural intervention for children with ASD. The architecture is organised with different modules to sense the environment and interaction, to make decisions based on the history of human-robot interaction over time, and to generate the robot's actions. This architecture has many merits but it has never been tested in a realistic, complex scenario. It could also be improved through support of non-reactive behaviours and behavioural adaptation to that of the young user. In another approach, Cao et al. propose a social behaviour control architecture capable of adapting to different therapeutic scenarios to achieve the goal of the interaction [34].

1.2 First steps towards Robot-Enhanced Therapy

As has been argued above, there is a need for the next generation of RAT – which we term *Robot-Enhanced Therapy* (RET) – to go beyond current WoZ paradigms. This approach is grounded in the ability to infer a child's psychological disposition and to assess their behaviour. The robot is then provided with the information necessary to select its next actions within well-defined constraints under supervision of a therapist. The latter aspect is key, as from an ethical perspective, there are strong indications that a fully autonomous system is not actually desirable in the context of interaction with vulnerable children [35, 36].

Consequently, RET robots should adopt a compatible, yet pragmatic approach concerning the desired level of autonomy. This entails restricting the modelling of psychological disposition to relatively simple emotions, immediate intentions, and goals, and assessing the child's behaviour based on cues given through body movement, facial expression, and vocal intonation. This will allow the robot to react to the child's requests in a contingent manner, to record, and to give specific feedback to the child. All elements would be conducted in a manner consistent with the level of attention and competence of the child. Such RET would not be entirely unlike Animal-Assisted Therapy (AAT), but possesses the added benefit that the robot can be instructed to behave in a specific manner and can be programmed to recognise situations where the therapist must resume control of the therapeutic intervention. The robot's autonomy therefore remains supervised in the sense that the therapist provides either high-level instructions for the robot or is called upon by the robot to interpret situations or data which it cannot reliably interpret itself. Thus, the aim of RET is not to replace the therapist but rather to provide them with an effective tool and mediator, embedded in a smart environment of which they remain in full control.

There are some additional desiderata for RET. For example, since RET will be an applied field where a number of therapists might be working independently, it is desirable to ensure that robot controllers developed for such an application be as platform-independent as possible. Also, children require therapy tailored to their individual needs, and RET robots must be able provide this. To achieve this, research in a clinical framework should investigate how children with ASD behave and perform during interactions with a therapeutic robot compared to a human partner, with respect to different social behaviours.

The EC-FP7 funded DREAM project [37] (Development of Robot-Enhanced therapy for children with AutisM spectrum disorders) is making progress in this direction. The aim is to reduce the workload of the therapist by letting parts of the intervention be taken over by the robot. This includes, for example, monitoring and recording the behaviour of the child, engaging the child when they are disinterested, and adapting between levels of intervention. This enables the therapist to oversee different children and plan the required intervention for every child on an individual basis.

The purpose of this paper is to highlight the steps completed in developing DREAM, the first robot-enhanced therapy project. Section 2 describes the clinical context where the project is to be tested, defines the measured variables, children and environmental conditions, and reveals first results. We deepen the concept of supervised autonomy in Section 3, detailing the control architecture.

Finally, we conclude with a synthesis of the lessons learned and take-home messages in Section 4.

2 Clinical framework

In order to evaluate the effectiveness of RET robots for improving social skills in children with ASD, several specific behaviours were observed during therapy sessions. These include: reciprocal turn-taking, shared attention, social reciprocity, sustained interaction, eye-gaze, spontaneous interaction, imitation of novel acts, and more. These behaviours are core problems in autism, representing both potential pathogenetic mechanisms and clinical symptoms/signs (e.g., deficit in social communication). In particular, we primarily target the following behaviours: imitation, turn taking, and joint attention, because we consider these to be the mechanisms that underlie other clinical symptoms, such as social and communication deficits.

From a clinical point of view, we aim to teach the aforementioned behaviours during repeated sessions of interactive games using social robots. This training is expected to lay a foundation for developing a set of implicit rules about communication; rules that will be transferred to interactions with human agents. The clinical goal behind this project is to determine the degree to which RET can improve joint attention, imitation and turn-taking skills, and whether or not this type of intervention provides similar, or greater, gains compared to standard interventions. For this purpose, the project was divided into two phases. During the first phase we employed RAT robots under a WoZ system, while in the second phase we will employ RET using a supervised autonomous system. The results from the first phase can be used as a baseline to compare the results of the second phase. Both phases will be compared to Standard Human Treatment (SHT) conditions.

In order to assess the effectiveness of both RET for children with ASD, we use single case experiments, more specifically, classic single-case alternative treatment design. In both RET and SHT conditions, children have 6 to 8 baseline sessions (in which we measure their initial performance on the variables under investigation), 8 standard therapy sessions and 8 RET sessions. Children participate in SHT sessions and RET sessions in a randomised manner to avoid ordering effects. In order to confirm children's diagnosis of autism and to assess their social and communication abilities we have used the ADOS instrument [38]. Apart from using ADOS as a diagnosis instrument we also use it as a measurement tool, in order to quantify differences in the obtained scores before and after interventions. After the initial ADOS application and baseline sessions considering the therapeutic setting, children interact directly with either a robot or human, with another person mediating the interaction between the child and the interaction partner in either condition.

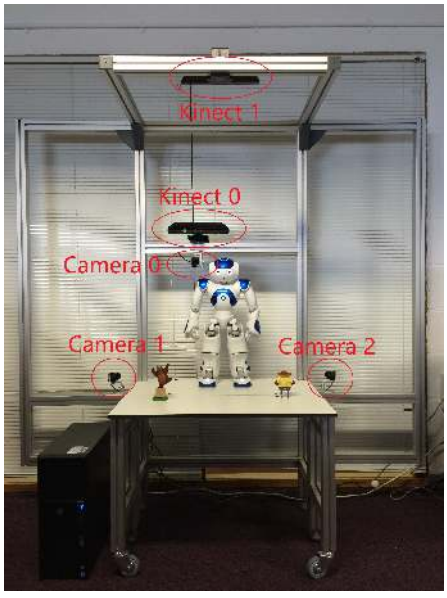


Figure 1: The intervention table and location of different cameras and Kinects.

All three tasks to be tested are implemented following the discrete trial format, a commonly used approach in early intervention programs for autism [39]. The elements that characterise this approach are: the teaching environment is highly structured; behaviours are broken into discrete sub-skills, which are presented over multiple, successive trials; and the child is taught to respond to a partner's discriminative stimulus (e.g., "Do like me!") through explicit prompting, prompt fading and contingent reinforcement [39].

To test our hypothesis, we use the humanoid robot NAO which acts as a social partner in each task, initiating behaviours like arm movements (for imitation purposes), taking turns and triggering joint attention episodes. An additional technological tool integrated in this research is the electronic "Sandtray" [40]. The platform uses a touchscreen and allows for social engagement through a collaborative interaction platform. The hardware consists of a 26-inch capacitive touchscreen and an associated

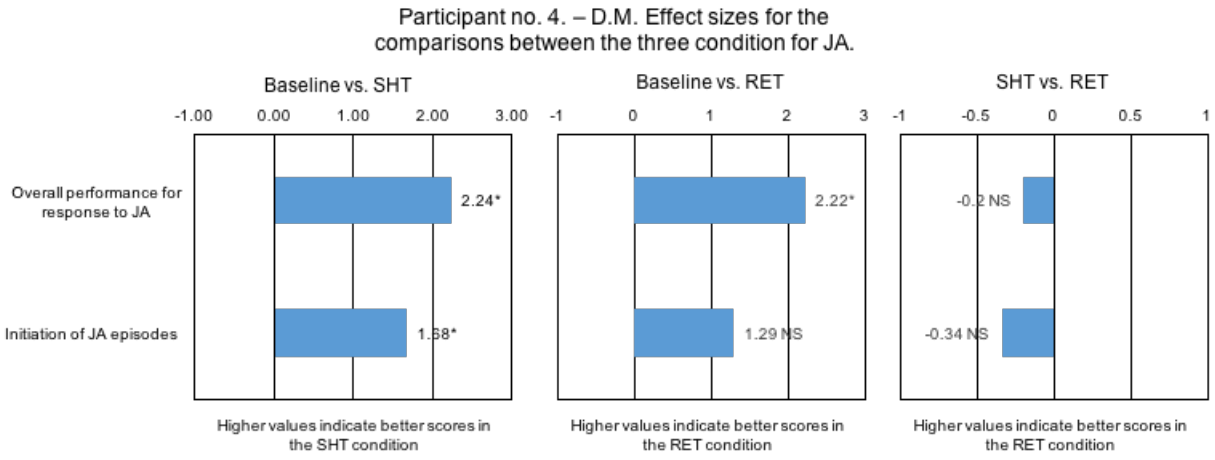


Figure 2: Cohen's *d* effect sizes for each comparison between the baseline, RET and SHT conditions. "*" indicates a statistically significant difference. "NS" indicates a comparison that is not statistically significant.

control server, upon which a series of pictures can be manipulated by dragging (on the part of the human partner), or simulated dragging (on the part of the robot partner).

In order to capture sensory information, we use an intervention table (shown in Figure 1), which accommodates ASD children interacting with the robot NAO. It employs five individual sensors, including three RGB cameras and two Microsoft Kinects, to capture the data. The cameras are used for face detection and gaze estimation. One Kinect has two functions: to capture both RGB and depth images for 3D facial feature extraction and 3D gaze estimation, and to detect skeleton joints for action recognition and hand tracking. The second Kinect is used to capture both RGB and depth images for robot and objects detection and tracking, see Section 3.1 for additional details. In order to keep the environment as close as possible to the standard intervention setting, we have used a small table and small chairs, also the distance between the robot and the child or between the therapist and the child was about 30 centimetres.

To assess the children's performance in the task, we measure two types of variables: primary and secondary. Primary variables comprise the performance of the child on the three tasks, based on task solving accuracy (e.g., movement accuracy in the imitation task, following instructions to wait for his/her turn on the turn taking task, and gazing in the joint attention task). Secondary variables involve outcomes such as:

a) social engagement: eye-contact and verbal utterances;
b) emotional level: positive and negative emotions;
c) behavioural level: stereotypical behaviours, adaptive and maladaptive behaviours;
d) cognitive level: rational and irrational beliefs.

For each of the measured variables, we provide an operational definition to be used as a basis for the learning process that maps child behaviours. This set of behaviours describes the child's actions during the intervention tasks in perceptual terms. This will provide the robot with the necessary input to react congruently and autonomously towards the child. Most of the variables are to be measured in frequency (e.g., eye contact – how many times the child looked at the upper part of the robot) except the beliefs where we would analyse the speech of the child and decide whether the phrase implies a rational or irrational statement (according to the definition of rational statement used in cognitive therapy).

Although several studies have been conducted in this field, our studies use a rigorous methodology, utilising an evidence-based paradigm, leading to the use of standard designs that involve several baseline measurements (e.g., single-case alternative treatments design), standard instruments for diagnosis (e.g., ADOS), and structuring the tasks developed based on empirical validated intervention techniques (e.g., discrete trial).

We now present the results obtained after completing the first phase of the project. Overall, the results of the experiments conducted in the WoZ paradigm show mixed results for the efficacy of RET, especially for primary outcomes (task performance, based on solving accuracy). The results differ from one task to another, such that in the turn-taking task RET seems to be as good as or even

better than SHT, especially for children with lower levels of prior skills. This means that some of the participants exhibit better performance when interacting with the robot compared to standard treatment. Regarding joint attention, the children's performance was similar in both conditions for the majority of the participants. However, for the imitation task, RET seems less effective than SHT. These results are important because they can help us to understand the conditions under which robots can be implemented in ASD therapy, and where the human therapist should be the main actor.

In the case of secondary variables, some differences are observed. In the imitation task, children looked more at the robot compared to the human partner, meaning that the children were interested in the robot partner during the entire intervention period. Regarding the emotional level, positive emotions appeared more in the imitation and joint attention tasks, where the robot was the interaction partner. As for the behavioural level, the presence of the robot in the task acts as a behavioural activator, so that both adaptive and maladaptive behaviours seem to appear more often in the RET condition compared to SHT condition (Figure 2).

The outcomes of these studies can serve as a basis for developing computational models capable of detecting inter-patient differences as well as tracking individual progress throughout the therapy. These models represent the foundation for developing a cognitive architecture with supervised autonomy (Section 3.3).

3 Supervised Autonomy

Effective child-robot social interactions in supervised autonomy RET requires the robot to be able to infer the psychological disposition of the child and use it to select actions appropriate to the current state of the interaction. How does the child feel? Are they happy, sad, disinterested or frustrated? Do they pay attention to the robot? What does their body language communicate and what are their expectations? Will they get bored in the therapy? The disposition can be inferred from gaze behaviours, body behaviours, and speech behaviours, see Section 3.1. Another important consideration is so-called "testing behaviour", which is described as a systematic variation of activity of the child while closely watching the other partner. This is related to perceiving intentions of others and to dynamics of imitation: role-reversal behaviours, turn taking, initiation of new behaviours, etc. Research towards supervised autonomy must develop computational models that can assess the behaviour of a child and infer their psychological disposition (Section 3.2). As noted already, we view these goals as a more pragmatic and less ambitious version of the well-known Theory of Mind problem, a problem for which a complete solution is not a realistic proposition in the near future.

The core of supervised autonomy, as described above, is a cognitive model which interprets sensory data (e.g., body movement and facial expression cues), uses these percepts to assess the child's behaviour by learning to map them to therapist-specified behavioural classes, and learns to map these child behaviours to appropriate therapist-specified robot actions. Thus, the DREAM system architecture has three major functional subsystems:

1. Sensing and Interpretation,
2. Child Behaviour Classification, and
3. Social Cognitive Controller.

The functional specifications of these three subsystems are derived from the different types of intervention targeted in Section 2. These interventions are described as a sequence of actions, each action comprising a number of constituent movements and sensory cues linked to a particular sensory-motor process. The motor aspect of these processes provides the basis for the robot behaviour specification to be implemented in the social cognitive control subsystem. The sensory aspect provides the basis for the sensory interpretation subsystems and also the child behaviour classification subsystem. The functional specification of the three subsystem components are described in detail in Sections 3.1, 3.2, and 3.3, respectively. The overall system architecture is shown in Figure 3.

In addition to the three subsystem components identified above, there is a Graphical User Interface (GUI) component to facilitate external control of the robot by a user (either a therapist or a software developer) and to provide the user with an easy-to-understand view on the current status of the robot control (Figure 4). It also provides a graphic rendering of the child's behavioural state, degree of engagement, and degree of performance in the current intervention.

3.1 Sensing and interpretation

In pursuing the goal of multi-sensory data fusion, analysis, and interpretation, RET cognitive controllers should target the features that are required by the three scenarios described in Section 2 (joint attention, imitation, and turn-

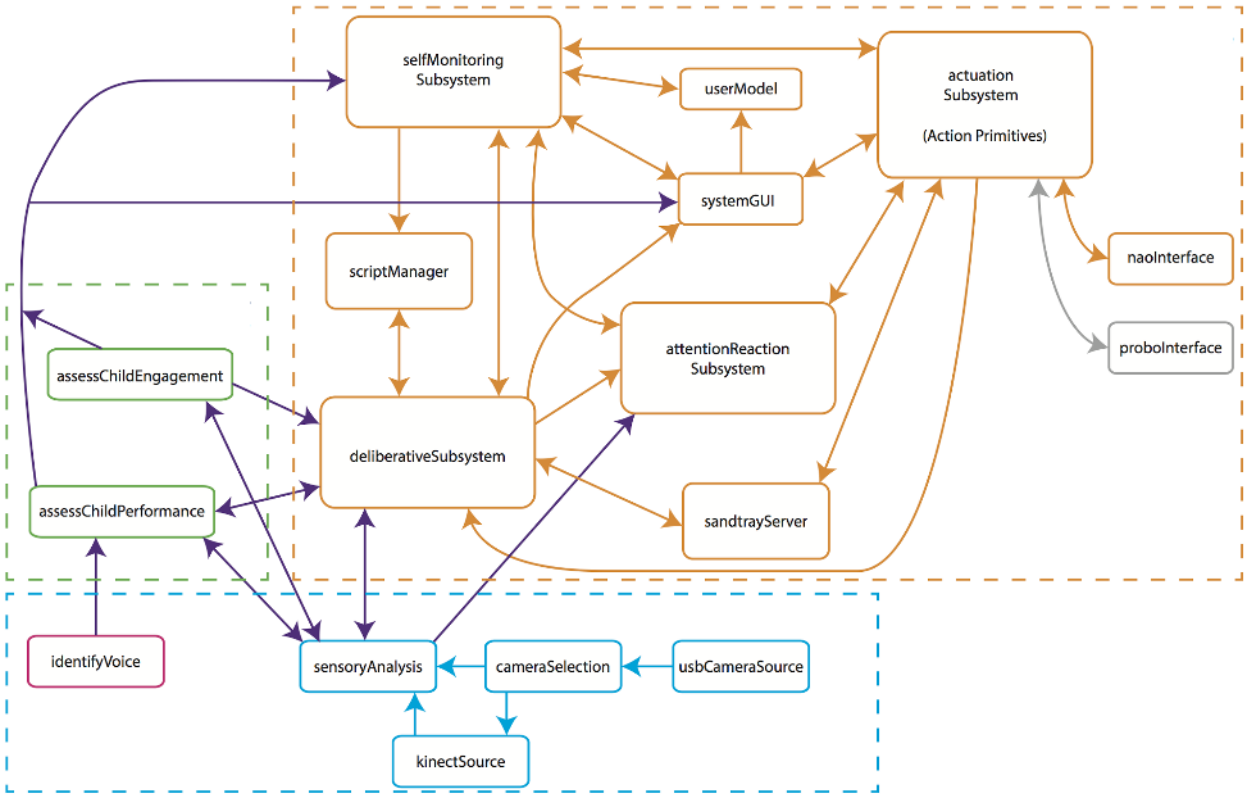


Figure 3: Graphical illustration of the DREAM software architecture. Arrows represent communication between components through one or more ports. Port names have been omitted for clarity. The three subsystems shown are: sensory interpretation (blue), child behaviour classification (green), and social cognitive controller (orange).

taking), at different levels of interaction complexity. These include:

1 Gaze analysis, including frequency and time of fixation on different parts of the robot, on other agents, on objects that are in front of the robot (for joint attention behaviours), and on faces in the peripheral field of view during a social interaction or play.
2 Frequency and duration of movements (the distance between the child and the robot, the position of the child in the space, interaction gestures, contact between objects and parts of the robot, and level of general activity, i.e., how much the child moves).
3 Vocal prosody, to identify statistics on congruent prosodic utterances between the child and the robot, such as the number of words, the frequency of verbal initiations and the length of the verbal speech during an interaction sessions (individuals with autism have difficulties in recognising and producing prosody and intonation [41]) and speech recognition in order to respond to the children's responses during the scenarios.
4 Vocal analysis of early speech measurement for key acoustic parameters [42].
5 Emotional appearance cues, in order to make explicit the dynamic processes that create, and are created by, the relationships with others [43].
6 Stereotypical behaviours, including the level of behavioural repetitiveness (such as shaking head, waving hand).

Multi-sensory data is used to provide quantitative support for the diagnosis and care/treatment of ASD children. This section shows the conclusions obtained after investigating methods and solutions for multi-sensory data perception and interpretation, with a focus on the complexity of extracting meaningful information about the ASD children. Specifically, techniques of gaze estimation, skeleton joint-based action recognition, object tracking, face and facial expression recognition, and audio data processing are presented.

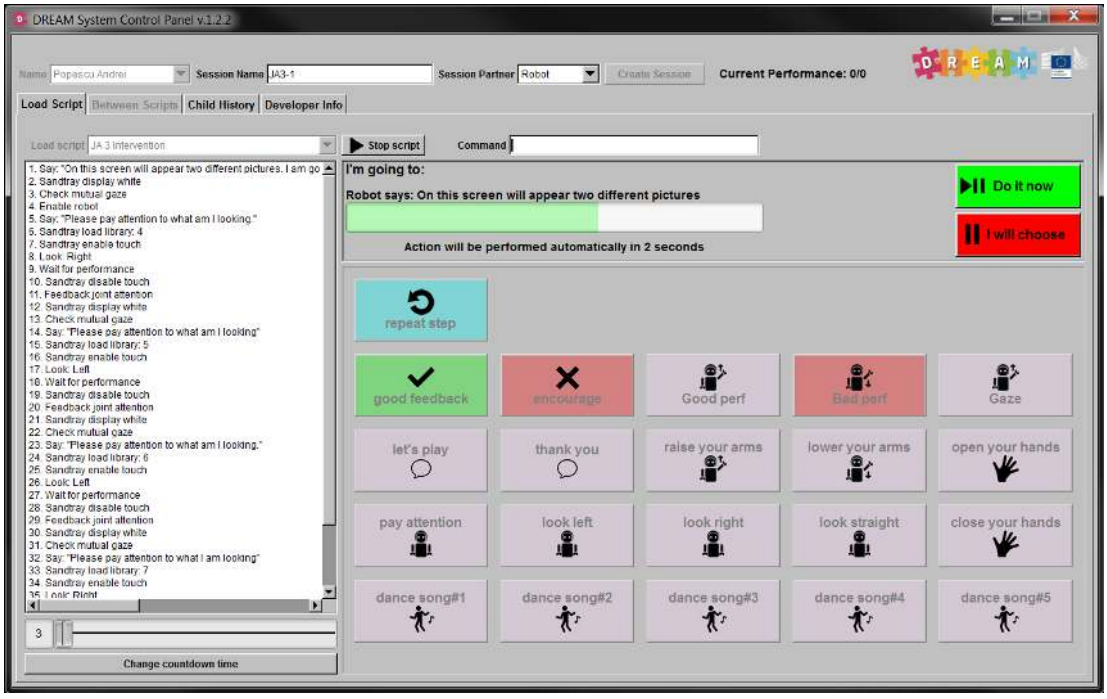


Figure 4: Graphical User Interface (GUI) component used by the therapist to control the robot. On the *left* side, the script of the intervention protocol is listed. On the *right* side, a set of robot actions available to overrule the autonomous behaviour.

3.1.1 Gaze estimation

The main challenges of gaze estimation involved in RET for ASD children are large head movement, illumination variation and eyelid occlusion. Although the designed multi-sensor system can successfully capture the child's face with large head movement, it is also a challenge to determine which camera can obtain the best view of the frontal face. To remedy this, we have proposed a multi-sensor selection strategy to adaptively select the optimal camera, see [44]. In the proposed strategy, all sensors are calibrated and used to capture the sensory data in parallel. In order to perform optimal camera selection, a face confidence score of each camera is defined. This score is acquired by measuring the variation of facial landmarks of a detected face with respect to facial landmarks of a predefined frontal face. The camera with the highest face confidence score will be selected as the optimal camera.

Once the face is detected, a Supervised Descent Method (SDM) trained with a database as described in [45] is employed to locate the feature points in the face and an object pose estimation method (POSIT) [46] is utilised to calculate the head pose. Then we propose an improved convolution based integro-differential method to localise the iris centres of the child [47, 48]. Compared with the conventional integro-differential method [49], the improved method is computationally much faster and it also achieves higher accuracy even in challenging cases of partial eyelid occlusion occurs or illumination varies (as shown in Figure 5).

Based on the obtained head pose and iris centres, we have proposed a two-eye model based method to estimate the final point of gaze of the ASD child. The proposed method averages the gazes of both eyes for a final gaze estimation. Moreover, we calculate the personal eye parameters by approximating the visual axis as a line from the iris centre to the gaze point. Experimental results show good performance of the proposed gaze estimation method (as in Figure 6).

3.1.2 Human action recognition

In the intervention task of imitation, either the ASD child's actions or the therapist's actions should be recognised when the child interacts either with the therapist or with the robot. Early proposed approaches mainly recognise human action from 2D sequences captured by RGB cameras [50–52]. However, the sensitivity to illumination changes and subject texture variations often degrades the recognition accuracy. These problems can be solved by using depth information acquired from a depth sensor

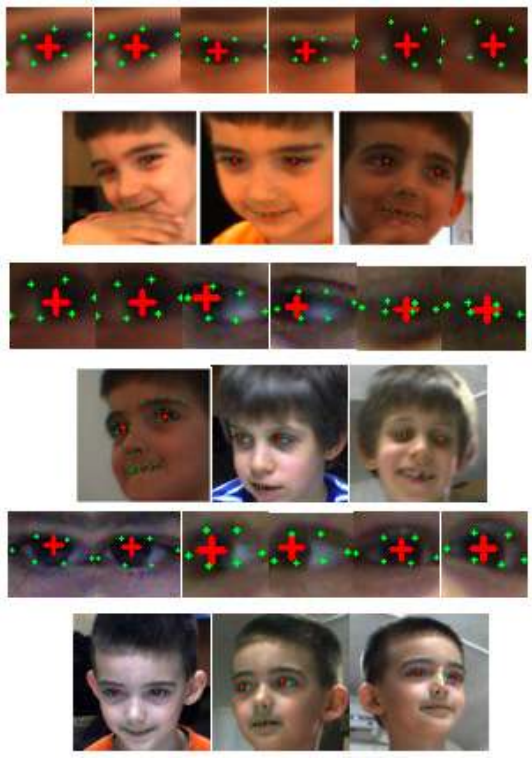
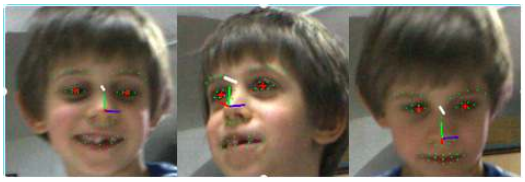


Figure 5: Images obtained from the intervention table localising the iris centres of the child.

(a)

(b)



(c)

Figure 6: Gaze estimation results on an ASD child recorded from intervention table. The white line denotes the gaze direction. (a) Estimation with camera0. (b) Estimation with camera1. (c) Estimation with camera2.

since images from depth channel can provide another dimensional information.

The main idea is to represent the movement of the body using the pairwise relative positions of the skeleton joint features that can be extracted by a Kinect. We have utilised the Kinect SDK for acquiring the skeleton data (as shown in Figure 7). For each child, ten joint positions are tracked by the skeleton tracker. The position coordinates are then normalised so that the motion is invariant to the initial body orientation and the body size. We have also presented a novel skeleton joint descriptor based on 3D Moving Trend and Geometry (3DMTG) property for human action recognition, see [53]. Specifically, a histogram of 3D moving directions between consecutive frames for each joint is constructed to represent the 3D moving trend feature in the spatial domain. The geometry information of joints in each frame is modelled by the relative motion with the initial status. After creating the descriptor, a linear Support Vector Machine (SVM) classification algorithm [54] is used for action recognition. We have evaluated the proposed method on a publicly available dataset MSR-Action3D [55] and the results demonstrate that our method can achieve high recognition rates on both similar actions and complex actions.

Based on the detected skeletal joints, we can also track the 3D position of a hand, frame by frame. Hand tracking can assist in estimating the location of object to grasp and is a key step for gesture recognition [56]. This will be used to analyse which object is grasped by the ASD child and to help with the activity classification.

Figure 7: Skeleton joints detection for ASD children.

3.1.3 Facial expression recognition

We have used the Local Binary Patterns feature extraction method on Three Orthogonal Planes (LBP-TOP) to represent facial appearance cues and applied the SVM for identity and facial expression classification [57].

Local Binary Patterns (LBP) is a non-parametric method and has proven to be a powerful descriptor in representing the local textural structure [58]. The main advantages of LBP are the strong tolerance against illumination variations and the computational simplicity. This method has successfully been used in both spatial and spatio-temporal domains in face recognition and facial expression recognition.

The LBP-TOP has been validated as effective for facial expression recognition as well as dynamic texture analysis, see [59]. The challenges in LBP-TOP are face registration and identity bias. LBP-TOP needs each frame in an image sequence to be in the same size, or at least the subregions of each frame to be in the same size. Any in-plane or out-plane rotation will degrade its performance. An effective LBP-TOP operator is highly dependent on face registration. The problem of identity bias generally exists in low-level features, which means that the extracted features reserve more information about identity rather than expressions.

To solve the above mentioned problems, we have proposed an approach to automatically recognise emotions using local patch extraction and LBP-TOP representation. We first detect point-based facial landmark by means of SDM and then extract local patches according to fiducial points. By doing so, the effect of identity bias can be better mitigated since the regions around fiducial points preserve more expression-related cues. Moreover, within all the frames in a sequence, the location of subjects (e.g., eyes, nose) are more stable and facial texture movements are more smooth. In each patch of sequence, block-based approach is exploited where LBP-TOP features are extracted in each block and connected to represent facial motions.

3.1.4 Object tracking

Numerous object detection and tracking algorithms have been proposed in the literature. This functionality is necessary to detect and track the objects (toys) on the intervention table and finally to judge whether the objects are picked up by an ASD child or not. The main challenges are object variety, illumination and occlusion. To effectively detect and track objects in real time, a blob based Otsu object detection method [60] is firstly employed to detect the objects. Then the GM-PHD tracker [61] is employed to track the objects over time due to its good performance in multi-object tracking. In the object detection stage, we have used the Otsu algorithm for adaptively image binarisation and employed the blob algorithm to detect the regions of the objects. The centre of each blob is regarded as the position of each object.

Object detection can find all the locations of objects on the table at each frame. To correctly associate the objects in consecutive frames, an efficient GM-PHD tracker is utilised for object tracking. In the object tracking stage, we have utilised an entropy distribution based method [62] to estimate the birth intensity of the new objects. Moreover, we have handled the partial occlusion caused by hand grasping based on a game theoretical method [63]. By doing so, objects in consecutive frames can be successfully and accurately tracked with correct identities. Figure 8 shows the results of object detection and tracking when a ASD child is interacting with a therapist. The results illustrate that our method can successfully detect and track objects even when they are occluded by hands. To obtain the 3D locations of the objects, a 2D-3D correspondence method [64] according to the depth information captured by the Kinect has been incorporated.

Figure 8: Object detection and tracking results.

3.1.5 Audio processing

The audio processing in RET must include speech recognition, sound direction recognition and voice identification. The speech recognition method is based on Microsoft Kinect SDK. We have utilised the trained model provided by the SDK to recognise the speech. To make the speech recognition individually independent, a dictionary is designed to store the predefined key words and related short sentences. The dictionary is fully customisable, which provides the convenience of recognising what sentences the subject has said by key words. The system starts to recognise the speech and

returns a textual representation on the screen when the subject speaks.

The direction of a sound is identified based on the different locations of microphones in the Kinect. Generally, the sound arrives at each of the microphones in a chronological order as the distances are different between microphones and the sound source [65, 66]. A signal with higher-quality sound will be produced by processing the audio signals of all microphones after calculating the source and position of the sound. Two significant properties, which are the sound angle and the confidence of the sound angle, are identified and then the system outputs the direction of the most crucial sound. We use a confidence score to represent the strength of the sound from the output direction. The larger the score is, the more confidence in accurately locating the sound.

Identity recognition remains a critical premise of autonomous perception in diagnostic support aimed at children with ASD. Among various off-body sensory modules for identity recognition, voice identification differentiates the subjects according to their acoustic data, which provides reliable identification without suffering from constraints of varying posture or behaviour. The identity logs of the child and the therapist are checked against the task specification and expectation, so that the response order matching or mis-matching will be further used for evaluation and diagnosis. Classifiers like Gaussian Mixture Model (GMM) and Vector Quantification (VQ) in combination with Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding (LPC) features are adopted in this project [67] to label the voice signal generated by the therapist and children with ASD.

3.1.6 Remaining challenges in sensing and interpretation

The proposed methods from Sections 3.1.1 to 3.1.5 are not without limitations. Below we describe some practical challenges which do not currently inhibit the performance of the therapy but would ideally be solved in future developments:

- Regarding the methods developed for gaze estimation, subjects are required to face the intervention table, described in Section 2, within the ranges of 120 degrees vertically.
- For human action recognition mechanisms in Section 3.1.2, large-scale body overlap would cause error in body joints tracking, and further lead to inaccurate human action recognition.
- In the case of facial expression recognition, large head post causing face distortion would influence the facial expression recognition accuracy. Moreover, face expression recognition works better for 'happy' detection compared to others, due to similarities in facial appearances for these expressions.
- The integrated object-tracking algorithm is limited to track objects in the context of a clear background (i.e., a white table).
- For audio processing (Section 3.1.5), speech recognition only supports English, and sound direction is limited from -50 degrees to 50 degrees horizontally (this is an assumption about where the sound would be expected).

3.2 Child behaviour classification

To operate in a supervised-autonomy mode, it is necessary to appraise the current behaviour of the child. This brings together the strands previously discussed in Sections 2 and 3.1. This appraisal happens in two stages (Figure 9). In the first stage, the data collected from the sensory interpretation setup (Section 3.1) is mapped onto the behaviours identified as relevant by the therapists (Section 2). This mapping draws on process knowledge from therapists, used to create and annotate training and validation sets of example child-robot interactions. The outcome of this process is not a winner-takes-all; rather, the classifiers – here, we use support vector machines trained on trajectories of the child's skeleton joints (Section 3.1) – identify the probability that a given behaviour is currently observed, for all behaviours.

This set of behaviours and probabilities are fed into the second stage. Here, the system attempts to derive the child's level of engagement, motivation, and performance on the current task, based on the interaction history (as derived from the first stage classifications). This is a challenging task, drawing heavily on therapists' semantic interaction knowledge, which provides insights into expected patterns given certain levels of engagement, motivation, and performance.

At both stages, the classifiers can be understood as generating real-time annotations of a therapy session of a similar type that therapists would normally create by recording such a session and annotating the files using ratings from multiple therapists. It follows from this insight how classifiers can be validated: auto-generated annotation files from (recorded) sessions that function as training data can both be submitted to therapists for verification. They can also be compared to existing

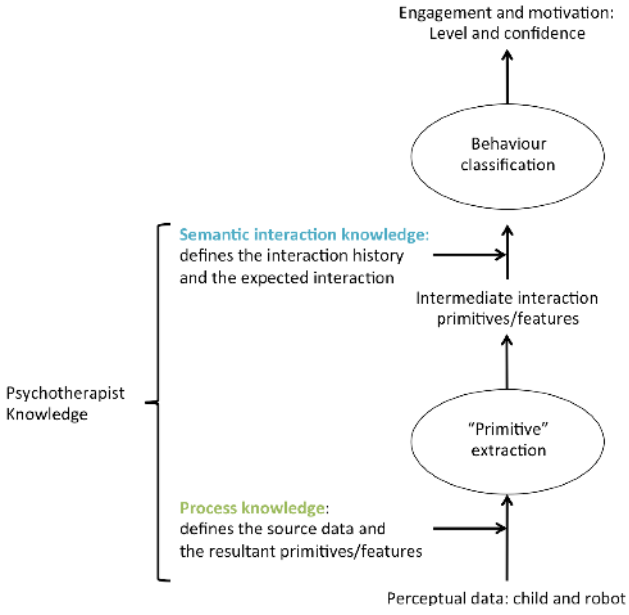


Figure 9: Child behaviour analysis flow.

annotations from therapists using standard inter-rater agreement measures.

Overall, it is worth noting that access to the therapists' knowledge is crucial for the success of this part of the work. It also clearly scopes the ambitions. There have been previous attempts at deriving general models of engagement (for a review, see [68]). However, we seek to build a system that operates to the specific requirements of the therapists.

The classifiers are explicitly allowed to report failures (in the sense that no defined behaviour could be defined and/or assessed). In any event, the outputs are fed into the cognitive controller of the robot (see next section), which decides future actions of the robot based on the classifier outputs (including the possibility that the classifiers failed to provide useful information). In addition to allowing supervised-autonomous operation of the robot, the developed classifiers offer other benefits:

- It allows a quantified evaluation of the evolution of a child's performance both within a single therapy session and over longer durations covering multiple sessions. Such quantifications might provide useful in future evaluation of therapeutic, as well as for assisting therapists in diagnostic tasks.
- The availability of such automated identification of psychological disposition can relieve therapists of some of their burden since it could be used, for instance, to automatically annotate videos of interactions with the children. To date, therapists are required to do this manually. As noted above, this reverse process forms, in fact, part of the validation exercise for the classifiers.

3.3 Social cognitive controller

Traditionally, cognition has been organised in three levels [69, 70]: the reactive, the deliberative and the reflective. In this section, we describe how these levels map onto our social cognitive controller.

The aim of the cognitive controller is to provide social robots with a behaviour underlying social interaction, which permits the robot to be used in RET in a supervised autonomous manner. This involves both autonomous behaviour and behaviour created in supervised autonomy, whereby an operator requests certain interventions, which are then autonomously executed by the robot. The cognitive controller is platform independent: rather than controlling actuators and modules specific for a robot platform, the cognitive controller sets parameters in descriptions and representations that are common across all platforms. This platform independence and high level representation of action allow this cognitive controller to operate with different robots in multiple therapy scenarios, see [34], entertaining or educating the child for limited periods.

The autonomous controller is composed of a number of subsystems which interact (Figure 10) and combine their suggested actions to produce a coherent robot behaviour, in the context of constraints laid down by the therapist (for example, the script to be followed, types of behaviour not permissible for this particular child because of individual sensitivities, etc). The cognitive controller architecture further defines the control that the supervising therapist can exert over the behaviour of the robot (effectively a limited 'remote control' functionality).

3.3.1 Socially reactive subsystem

The reactive level constitutes low-level processes which are genetically determined and not sensitive to learning in natural systems. This level is essential in social robots as it creates the illusion of the robot being alive, acting as a catalyst for acceptance [71]. The role that the reactive subsystem plays in generating the executed robot behaviour depends on the processing within the deliberative subsystem, and the oversight of the therapist (through the self-monitoring subsystem as interacted with through the system GUI). This means that, as with other

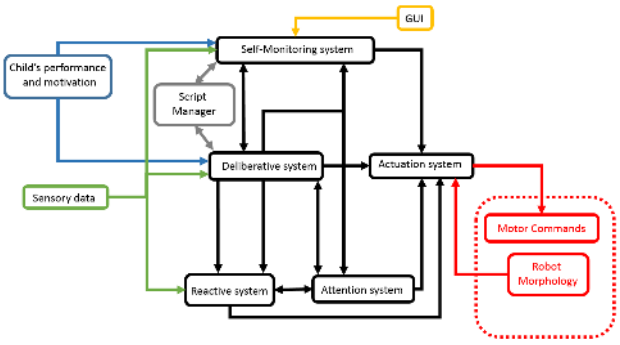


Figure 10: Description of the cognitive controller subsystems and how information flows from one subsystem to another.

layered control architectures (e.g., subsumption), the reactive subsystem contributes to, rather than completely specifies, the overall robot behaviour.

A general high level description of the reactive subsystem is shown in Figure 11. This describes how, given the sensory information and the inputs from the deliberative subsystem, the robot reacts to the current situation.

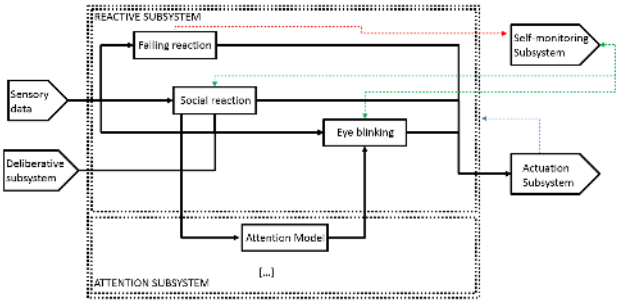


Figure 11: High level description of the reactive subsystem.

The reactive subsystem is composed of a number of modules as follows (see [72] for further details). Changes in balance may end up in a fall. In such cases, all active behaviours are interrupted, and a damage avoidance behaviour that fits the situation is triggered, see [73] for a case of minimising damage to a humanoid robot, and [74] for a case of a NAO robot that modifies its falling trajectory to avoid causing injuries in people in front of it.

In social situations, multiple verbal and non-verbal interactive encounters may occur. The child may or may not behave favourably towards the robot. These situations may be conflicting and special attention is required given the potential audience of this project. If it would be the case of a regular social robot, for both situations the robot may appropriately react, but under these circumstances, the reaction is simplified to facial expressions and speech acts, always under the supervision of the therapist.

The acceptability of the robot can be further increased if the robot mimics human blinking behaviour. Simulating blinking behaviour requires a human-level blinking model that should be derived from real human data. Several works have considered the dependencies of human eye blinking behaviour on different physiological and psychological factors. Ford et al. proposed the "blink model" for Human-Robot Interaction (HRI), which integrates blinking as a function of communicative behaviours [75]. For this reason, we adopt Ford et al.'s model to cover our needs and to provide accurate data for implementing the model.

Along with social reactions, the cognitive controller includes an attention subsystem to allow the robot to know the relevant stimulus in the scene [76]. This subsystem is a combination of perceptual attention, in which perceptual stimuli (reported by, for example, sound localisation; Section 3.1) that are particularly salient in the current context have to be selected, and attention emulation (from the deliberative subsystem) directs the robot's attention and gaze. These inputs provide the robot with a locus of attention that it can use to organise its behaviour.

Given the context in which this subsystem is implemented, attention behaviour has been divided between scripted (where the attention is determined by the requested scenario) and non-scripted interactions. Within scripted interactions, the highest priority is given to the deliberative subsystem outputs. Therefore, each time attention emulation is triggered, the point of interest is where the robot will look at, unless the therapist decides to override such behaviour.

Within non-scripted interactions, the attention model seeks the next point of interest to look at. For this purpose we have built a target selection algorithm adapted from [77] where the authors present a bottom-up attention model based on social features. Some simplifications of the model were applied to adapt it for our context. Other approaches like [78] were taken into account. This approach merges top-down and perceptual attention in an efficient manner. However, for the sake of simplicity we opted for adapting Zaraki et al.'s model due to implementation ease.

3.3.2 Deliberative subsystem

The deliberative subsystem is the primary locus of autonomous action selection in the cognitive controller

(Figure 10). This subsystem takes as input sensory data: child behaviour information, information on what step should be next executed from the therapy script, and higher-level direction from the therapist. It then proposes what action should be taken next by the robot. A central aspect of the cognitive controller is its ability to follow intervention scripts as defined by the clinicians for both diagnosis and therapy. These scripts describe the high-level desired behaviour of the robot, and the expected reactions and behaviours of the child, in a defined order. In a normal script execution context, the deliberative subsystem is the primary driver of behaviour, which would typically propose the next script step. There are however a number of circumstances in which this is not the most appropriate action to perform. For example, if the child is detected to have very low engagement with the task (as determined from the child behaviour analysis, and/or information from the sensory system saying the child is looking away for example), then it would be appropriate to attempt to re-engage the child with the robot/task prior to executing the next stage in the therapy script. In this case, the deliberative subsystem can choose to depart from the behaviour defined in the script, and instead propose a different behaviour.

The script manager itself, see Figure 10, separates the logic necessary to manage progression through the script (by taking into account the available sensory feedback after actions for example) from the script itself. This makes it straightforward to add new scripts or modify existing scripts as required. This logic management has in the first instance been achieved using a Finite State Machine (FSM).

There is currently no algorithm in the literature completing all the desiderata for our Action Selection Mechanism: keeping a supervisor in control whilst providing autonomy and adaptivity to the robot. Classical learning algorithms (such as classical Reinforcement Learning [79]) rely on exploration which could end with the robot executing actions that have a negative impact on the child. Algorithms such as Deep Learning [80] require large datasets to be able to learn (which do not currently exist for this application domain). An alternative for RET is to use the knowledge of the therapist to teach the robot appropriate actions using Interactive Machine Learning [81, 82] by allowing the human to provide input at run time to guide the robot action selection and learning.

Algorithms used in Interactive Machine Learning frameworks often only use the human to provide feedback on the robot actions to bootstrap the learning. Whilst allowing the robot to learn faster, these approaches do not use the human inputs to their maximum. We take stronger inspiration from the Learning from Demonstration community [83, 84] and give control of every action executed by the robot to the therapist. Following this approach, a new method was developed, termed SPARC (Supervised Progressively Autonomous Robot Competencies) [85, 86]. As shown in Figure 12, the goal of SPARC is to provide the robot with online learning, reducing the workload on the therapist whilst maintaining high performance throughout the interaction.

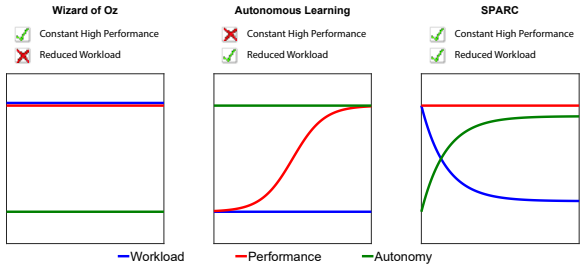


Figure 12: Comparison of expected ideal behaviours for three control approaches for RET on the robot's autonomy, robot performance, and workload on the therapist. The aim is to maintain high performance throughout the interaction while keeping the workload on the therapist as low as possible. By using Interactive Machine Learning and providing the therapist with control, SPARC is expected to meet these two key considerations.

SPARC relies on a suggestions/correction mechanism, by which the robot proposes actions to the supervisor who can passively accept the action or actively correct it. The resulting action is executed by the robot and the supervisor decision is fed back to the learning algorithm to improve the suggestion for the future (Figure 13). The states used for the learning are comprised of internal states of the robot and external states in the social and physical environment, including the child. Using the therapist's commands and correction, SPARC gradually builds up a state-action model, and as the interaction progresses, suggests more appropriate actions to the therapist.

SPARC is agnostic of the algorithm used; studies have been conducted using a neural network [85] and reinforcement learning [87] but there is no indication that it could not be used with other world representations or learning algorithms. In the first study, the results show that when the robot is learning, the workload on the supervisor is lower. This supports the idea that using learning algorithms to learn from a therapist controlling a robot in RET could lead to a reduction of workload. The therapist could subsequently focus more on the child behaviour, rather than having to focus only on controlling the robot.

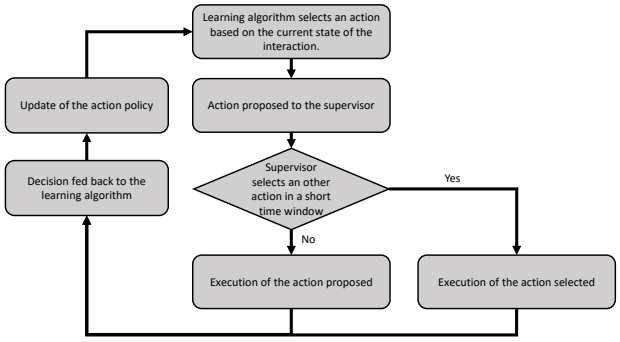


Figure 13: High-level action selection and learning flow used in SPARC.

The second study compared a SPARC based reinforcement learning to Interactive Reinforcement Learning [88], a more classical approach where rewards from the human are simply combined to environment rewards. Results have shown that SPARC allows faster and safer learning and that the control given to the supervisor prevents the robot from executing undesired actions whilst providing enough inputs to learn an efficient action policy.

3.3.3 Self-monitoring subsystem

As explained above, the social robot will always be under the supervision of a therapist or teacher. However, the controller should aim to act autonomously for as long as possible. A self-monitoring system plays the role of the reflexive level of the robot and has been designed as an alarm system [89]. An internal one is used when the robot detects that it cannot act because of a technical limitation or an ethical issue. An external alarm is one where the therapist overrules the robot behaviour selection.

This subsystem is always on and normally does nothing but monitor processes. When the alarm system switches on, an appropriate behaviour of the robot is initiated as it is undesired that the robot simply freezes its motions, which may look unnatural to the child. If an internal process creates the event, the robot switches to neutral interaction and asks for therapist help.

Through the reflexive level, the social cognitive controller manages possible ethical limitations. DREAM is concerned about the ethics of robotics and specifically, with how exactly the supervision or overruling will be implemented. Discussions include whether any overruling of the robot's behaviour by the therapist needs to be explicit (so that the child can understand that the behaviour of the robot is overruled by the therapist; it can also make errors just like any other social agent) or needs to be hidden (for instance, through previously defined code words, so the child does not recognise that the robot's behaviour is being modified).

The ethics of technology draws on fields in the social studies of science and technology and the philosophy and anthropology of technology [90, 91]. Moreover, in the last decade a specialised field entirely dedicated to ethics in machines and robots has grown out of philosophy [92].

We have conducted a survey [35] to understand the opinions of parents and therapists about social robots, and whether they believe robots can and should be used for ASD therapy for children, in order to inform roboticists, therapists, and policy makers about the ethical and social issues involved in RAT. One important finding in the survey was the positive acceptability of robots for helping children with autism compared with the negative feedback given in the Eurobarometer [93]. The survey included responses from parents of children with ASD (22%), and therapists or teachers of children with ASD (16%), the rest of the cohort was made up of students of psychology or people involved in organisations. Questions presented to the stakeholders were wide-ranging and included the following "Is it ethically acceptable that social robots are used in therapy for children with autism?" Of which the majority of interview respondents agree (48%) and strongly agree (37%). "Is it ethically acceptable to use social robots that replace therapists for teaching skills to children with autism?" With only 18% (agree) and 08% (strongly agree). This survey indicated the importance of stakeholder involvement in the process, focused around specific health care issues.

3.3.4 Platform independent flavour

The cognitive controller outputs the social actions of the robot, including non-verbal (facial and body) and verbal expressions. Such a controller needs to be independent of the robotic platform, as generic methods are required to control the robot's expressions, gestures and mobility. The goal of the actuation subsystem is to translate the actions of the social behaviour into readable social verbal and non-verbal cues, especially for our particular audience of young users with ASD. This subsystem determines which combination of low-level actions the robot should execute next, and how these actions are to be performed. Suggestions for actions to take come from the other subsystems. Along with this, it is assumed that the supervising therapist, through the GUI, will determine

(either beforehand or in real-time) the aspects of robot behaviour that should be executed, from which relative priorities will be determined for the three subsystems.

A number of robots capable of gesturing have been developed to study different aspects in HRI. Gestures implemented in robots are however, until now, subject to two important limitations. Firstly, the gestures implemented in a robot are always limited to a set of gestures necessary for the current research, and often limited to one type of gestures, see [94] for an example. The reason for this can be found in the second limitation: gestures are mostly preprogrammed off-line for the current robot configuration. The resulting postures are stored in a database and are replayed during interaction. This is the case for, among others, Robovie [95], HRP-2 [96] and Kobian [97]. Since the postures are dependent on the morphology, they cannot be used for other robots with other configurations. The result is that, when working with a new robot platform, new joint trajectories to reach the desired postures need to be implemented, which can be time consuming. It would however be much more efficient to make the implementation of gestures more flexible and to design a general method that allows easily implementing gestures in different robots.

Our method divides the robot embodiment in three areas: the face expression, developed to provide the behaviours with natural and emotional features; the overall pose, developed to calculate gestures whereby the position of the main parts of the body is crucial; and the end effector, developed for pointing and manipulation purposes.

Different robots use the Facial Action Coding System (FACS) by Ekman [98] to abstract away from the physical implementation of the robot face. FACS decomposes different human facial expressions in the activation of a series of Action Units (AU), which are the contraction or relaxation of one or more muscles. We have already implemented the FACS methodology in Probo to express emotions [99]. The NAO robot does not possess the facial expressibility that Probo has, as it has 0 DOF in the face and the only mechanism that it has to express facial gestures is through the change of colors in its eyes. For such reason, an eyebrows system that will help to understand better emotional expressions on NAO's face has been developed, see [100] for further details.

In a similar way, Body Action Units (BAU) have been defined together with a Body Action Coding System (BACS), where the different gestures are decomposed in the activation of BAUs. This system avoids pre-programming of robot-dependent body poses and actions, which is relevant since humans are able to recognise actions and emotions from point light displays (so without body shape) [101]. The physical actuation of AUs will depend on the morphology of the robot: a mapping will be needed between AUs and the degrees of freedom, and thus to the joints of the robot, this mapping will be specific to a robot platform. To ensure a realistic and readable overall posture, it is necessary to take into account the relative orientations of every joint complex the robot has in common with a human. A base human model was defined, and the target postures were quantitatively described by the orientation of the different joint complexes in the model using the BACS. While the Facial AUs are defined as a muscle or a muscle group, our BAUs are based on the human terms of motion. The units are grouped into different blocks, corresponding to one human joint complex, such as the shoulder or the wrist. These blocks can subsequently be grouped into three body parts, namely the head, body and arm, which we refer to as chains. In that way, a base human model was defined, consisting of four chains; the head, the body, the left arm and the right arm. Although the leg movements also contribute to the overall performance of the gesture, for a first validation of the method we decided to focus only on the upper body movements. This method has been successfully validated on the virtual model of different robots through a survey. See [102] for further details on the method and validation.

To calculate pointing and manipulation gestures, another strategy is used. In some situations, for example when reaching for an object, the position of the end-effector is important and specified by the user. For pointing towards an object, several end-effector poses are possible to achieve a pointing gesture to the specified target. In that case, an optimal pose of the end-effector is chosen, according to a cost-function minimising the deviation from a defined set of minimum posture angles. This specified end-effector pose then serves as input to calculated the corresponding joint angles, using the same inverse kinematics algorithm as used for the calculation of emotional expressions. Figure 14 shows the calculated end posture for a reaching gesture at $(34, -34, 38)$ for three different configurations. The first column shows the joint configuration, while the second column shows the calculated posture for that configuration. The desired end-effector position is visualised by a sphere. In the top row, a 9 DOF human arm is shown, consisting of a two DOF clavicle, 3 DOF shoulder, 1 DOF elbow and 3 DOF wrist (virtual model comes from the RocketBox libraries [103]). Configuration 2 shows the ASIMO robot [104]. As for the human model, the targeted end-effector position was reachable, and a suitable end posture could be calculated,

as shown in the second row. Configuration 3 is that of the NAO robot. NAO is considerably smaller than the previous models, and as a result, the maximum reachable distance is smaller. The desired position is located out of the range of the robot. Therefore, the pointing condition is activated, and a suitable posture for a pointing gesture towards the specified point is calculated. See [105] for further information.

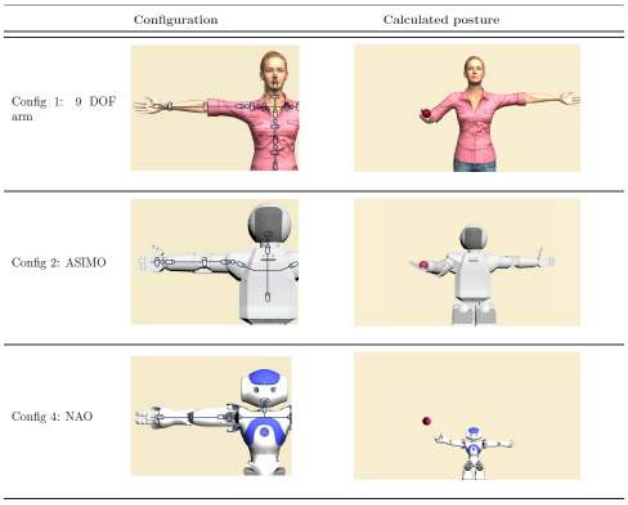


Figure 14: Results of the method for different arm configurations. The first column shows the joint configuration, while the second column shows the end posture for a place-at gesture at $(34, -34, 38)$.

4 Conclusion

Robot-Assisted Therapy is increasingly being used to improve social skills in children with ASD [106]. However, as discussed at the outset, there is a need for robots to move beyond the reliance on WoZ control of robots in therapeutic settings in a new paradigm that we term Robot-Enhanced Therapy. Section 1.1 discussed some of the challenges that researchers in RAT will face in these forthcoming developments. In particular, we highlighted the need for increasing the autonomy of the robot to improve therapeutic experiences.

To tackle these challenges, we recast them as practically solvable problems under a certain clinical framework in which therapeutic interventions are to be conducted. In Section 2, we described the measured variables and the clinical framework itself, providing us with a baseline to compare the performance of RET robots with RAT robots and to SHT conditions. Moreover, this framework functions as the starting point in the development of supervised autonomy systems.

As an insight into our first clinical study, we consider this work to provide a baseline to conduct second phase clinical studies with RET robots, although the results from this first phase showed mixed outcomes. There are still some limitations of using robots in clinical frameworks, such as delays due to the slow reaction time of the robot or connectivity problems between the robot and the therapist's computer. While we do not think they could have a strong impact on the performance of the child, they should be addressed in forthcoming projects. Overall, work such as that described here has the potential to impact clinical practices in therapy for children with ASD. The use of technology in the diagnosis process and interventions for individuals with ASD will ease the workload of the therapist and lead to more objective measurements of therapy outcomes.

Based on ethical studies concerning the acceptance of autonomous robots in therapies with children with autism, we suggest that a fully autonomous robot is not desirable, and aiming to achieve it is unrealistic. For this reason, a supervised autonomy approach is preferred. In Section 3, the supervised autonomy architecture is divided into three blocks: sensory information, child behaviour classification and social cognitive controller.

Sensory information collects, analyses and interprets data targeting the required features described in the clinical framework. We have successfully developed mechanisms for gaze estimation, human action recognition, facial expression recognition, object detection and tracking, speech recognition, voice identification and sound direction recognition, although constrained to specific application areas. These limitations are described in Section 3.1.6.

Realising that a full Theory of Mind is currently not realistic in RAT or RET scenarios, we reduced the problem to the identification of well-defined indicators of the child's level of engagement, motivation and performance on the current task. This classification is then used by the social cognitive controller, which allows the robot to act appropriately, given both its own autonomous behaviour, and behaviour defined by therapists. Given the conditions in which this architecture has been implemented, the robot behaviour has been divided between scripted and non-scripted interactions. Within scripted interactions, there is no room for the robot to be socially reactive and its behaviour is limited by the intervention protocol. However, the deliberative subsystem proposes actions to

the supervisor and learns from the therapist's choices building a state-action model. In non-scripted scenarios, the robot is being responsive to verbal and non-verbal interactive cues and suggests possible actions to re-engage the child in the intervention protocol. Robot actions must be expressed independently of the robotic platform therapists decide to use. Therefore, a platform independent method to implement these actions in robots with different sets of DOF is described.

To summarise, this paper described the insights gained from progress in the DREAM project so far, highlighting how the many elements involved in the solution of this complex problem come together. In particular, we have tackled some of the challenges underlying supervised autonomy in RET and described possible approaches to overcome them.

Acknowledgement: The work leading to these results has received funding from the European Commission 7th Framework Program as a part of the DREAM project, grant no. 611391. The Authors obtained a consent for the use of all the photos in this publication.

References

[1] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, Arlington: American Psychiatric Publishing, 2013.

[2] Howlin P, Goode S, Hutton J, Rutter M, Adult outcome for children with autism, Journal of Child Psychology and Psychiatry 45(2):212–229, 2004.

[3] Mordre M, Groholt B, Knudsen AK, Sponheim E, Mykletun A, Myhre AM, Is long-term prognosis for pervasive developmental disorder not otherwise specified different from prognosis for autistic disorder? findings from a 30-year follow-up study, Journal of autism and developmental disorders 42(6):920–928, 2012.

[4] Dawson G, Osterling J, Early intervention in autism, The effectiveness of early intervention, 307–326, 1997.

[5] Roberts JM, Ridley G, Review of the research to identify the most effective models of best practice in the management of children with autism spectrum disorders, Centre for Development Disability Studies, 2004, University of Sydney.

[6] Eldevik S, Hastings RP, Hughes JC, Jahr E, Eikeseth S, Cross S, Meta-analysis of early intensive behavioral intervention for children with autism, Journal of Clinical Child & Adolescent Psychology 38(3):439–450, 2009.

[7] Peters-Scheffer, N., Didden, R., Korzilius, H., and Sturmey, P., A meta-analytic study on the effectiveness of comprehensive ABA-based early intervention programs for children with autism spectrum disorders. Research in Autism Spectrum Disorders, 5(1):60–69, 2011.

[8] De Rivera C, The use of intensive behavioural intervention for children with autism, Journal on developmental disabilities 14(2):1–15, 2008.

[9] Ozonoff S, Reliability and validity of the Wisconsin Card Sorting Test in studies of autism, Neuropsychology 9(4):491, 1995.

[10] Diehl JJ, Schmitt LM, Villano M, Crowell CR, The clinical use of robots for individuals with autism spectrum disorders: A critical review, Research in autism spectrum disorders 6(1):249–262, 2012.

[11] Robins B, Dautenhahn K, Dubowski J, Does appearance matter in the interaction of children with autism with a humanoid robot?, Interaction Studies 7(3):509–542, 2006.

[12] David D, Matu SA, David OA, Robot-based psychotherapy: Concepts development, state of the art, and new directions, International Journal of Cognitive Therapy 7(2):192–210, 2014.

[13] Barakova EI, Lourens T, Expressing and interpreting emotional movements in social games with robots, Personal and ubiquitous computing 14(5):457–467, 2010.

[14] Chevalier P, Isableu B, Martin JC, and Tapus A, Individuals with autism: Analysis of the first interaction with nao robot based on their proprioceptive and kinematic profiles, In Advances in robot design and intelligent control, 225–233, 2016.

[15] Tapus A, Tapus C, Matarić MJ, User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy, Intelligent Service Robotics 1(2):169–183, 2008.

[16] Albo-Canals J, Yanez C, Barco A, Bahón CA, Heerink M, Modelling social skills and problem solving strategies used by children with ASD through cloud connected social robots as data loggers: first modelling approach, Proceedings of New Friends 2015: the 1st international conference on social robots in therapy and education, 22-23, 2016.

[17] Boccanfuso L, Scarborough S, Abramson RK, Hall AV, Wright HH, O'Kane JM, A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned, Autonomous Robots, 1-19, 2016.

[18] Yun SS, Kim H, Choi J, Park SK, A robot-assisted behavioral intervention system for children with autism spectrum disorders, Robotics and Autonomous Systems 76:58-67, 2016.

[19] Vanderborght B, Simut R, Saldien J, Pop C, Rusu AS, Pintea S, Lefeber D, David DO, Using the social robot probo as a social story telling agent for children with asd, Interaction Studies 13(3):348–372, 2012.

[20] Simut R, Costescu CA, Vanderfaeillie J, Van de Perre G, Vanderborght B, Lefeber D, Can you cure me? children with autism spectrum disorders playing a doctor game with a social robot, International Journal on School Health 3(3),(Inpress), 2016.

[21] Simut R, Pop C, Vanderfaeillie J, Lefeber D, Vanderborght B, Trends and future of social robots for asd therapies: potential and limits in interaction, presented at the International Conference on Innovative Technologies for Austism Spectrum Disorders (ASD): tools, trends and testimonials, 2012.

[22] Huijnen C, Lexis M, Jansens R, Witte LP, Mapping Robots to Therapy and Educational Objectives for Children with Autism Spectrum Disorder, Journal of autism and developmental disorders 46(6):2100-2114, 2016.

[23] Landauer TK, Psychology as a mother of invention.,ACM SIGCHI Bulletin 18(4):333–335, 1987.

[24] Wilson J, Rosenberg D, Rapid prototyping for user interface design, Handbook of Human-Computer Interaction 39:859–873, 1988.
[25] Scassellati B, Admoni H, Mataric M, Robots for use in autism research, Annual review of biomedical engineering 14:275–294, 2012.
[26] Thill S, Pop CA, Belpaeme T, Ziemke T, Vanderborght B, Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook, Paladyn, Journal of Behavioral Robotics 3(4):209–217, 2012.
[27] Cabibihan JJ, Javed H, Ang Jr M and Aljunied SM Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. International journal of social robotics 5(4):593–618, 2013.
[28] Robins B, Otero N, Ferrari E and Dautenhahn K, Eliciting requirements for a robotic toy for children with autism—results from user panels, Proceedings of the 16th IEEE international symposium on robot and human interactive communication (RO-MAN),101–106, 2007.
[29] Ferrari E, Robins B and Dautenhahn K, Therapeutic and educational objectives in robot assisted play for children with autism, Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 108–114, 2009.
[30] Michaud F, Duquette A and Nadeau I, Characteristics of mobile robotic toys for children with pervasive developmental disorders, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 3:2938–2943, 2003.
[31] Dennett DC, The intentional stance, MIT press, 1989.
[32] Arkin RC, Homeostatic control for a mobile robot: Dynamic replanning in hazardous environments. Journal of Robotic Systems, 9(2):197–214, 1992.
[33] Feil-Seifer D, Mataric MJ, B3IA: A control architecture for autonomous robot-assisted behavior intervention for children with Autism Spectrum Disorders, Proceedings on the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 328–333, 2008.
[34] Cao HL, Gómez Esteban P, De Beir A, Simut R, Van De Perre G, Lefeber D, Vanderborght B, Toward a platform-independent social behavior architecture for multiple therapeutic scenarios, Proceedings of Conference New Friends, 3-32, 2015.
[35] Coeckelbergh M, Pop C, Simut R, Peca A, Pintea S, David D, Vanderborght B, A survey of expectations about the role of robots in robot-assisted therapy for children with asd: Ethical acceptability, trust, sociability, appearance, and attachment, Science and engineering ethics 22(1):47–65, 2016.
[36] Peca A, Robot enhanced therapy for children with autism disorders: Measuring ethical acceptability, IEEE Technology and Society Magazine 35(2):54–66, 2016.
[37] Dream project, 2016, http://www.dream2020.eu/
[38] Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, Pickles A and Rutter M, The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism, Journal of autism and developmental disorders 30(3):205–223, 2000.
[39] Ingersoll B, The social role of imitation in autism: Implications for the treatment of imitation deficits, Infants & Young Children 21(2):107–119, 2008.
[40] Baxter P, Wood R, Belpaeme T., A touchscreen-based sandtray to facilitate, mediate and contextualise human-robot social interaction, Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 105–106, 2012.
[41] Shriberg LD, Paul R, McSweeny JL, Klin A, Cohen DJ and Volkmar FR, Speech and prosody characteristics of adolescents and adults with high functioning autism and Asperger syndrome, Journal of Speech, Language, and Hearing Research, 44:1097–1115, 2001.
[42] Oller D, Niyogi P, Gray S, Richards J, Gilkerson J, Xu D, Yapanel U and Warren S, Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development, Proceedings of the National Academy of Sciences, 107(30):13354-–13359, 2010.
[43] Halberstadt AG, Dunsmore JC and Denham SA, Spinning the pinwheel, together: More thoughts on affective social competence, Social Development, 10:130—136, 2001.
[44] Cai H, Zhou X, Yu H, Liu H, Gaze estimation driven solution for interacting children with asd, Proceedings of 26th International Symposium on Micro-Nano Mechatronics and Human Science, 1–6, 2015.
[45] Xiong X, De la Torre F, Supervised descent method and its applications to face alignment, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 532–539, 2013.
[46] Dementhon DF, Davis LS, Model-based object pose in 25 lines of code, International Journal of Computer Vision 15:123–141, 1995.
[47] Cai H, Liu B, Zhang J, Chen S, Liu H, Visual focus of attention estimation using eye center localization, IEEE Systems Journal 99:1–6, 2015.
[48] Cai H, Yu H, Yao C, Chen S, Liu H, Convolution-based means of gradient for fast eye centre localization, Proceedings of International Conference on Machine Learning and Cybernetics, 759–764, 2015.
[49] Timm F, Barth E, Accurate eye center localisation by means of gradients. Proceedings of 6th International Conference on Computer Vision Theory and Applications, 125–130, 2011.
[50] Bobick AF, Davis JW, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23:257–267, 2001.
[51] Niebles JC, Li F, A hierarchical model of shape and appearance for human action classification, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.
[52] Niebles JC, Wang H, Li F, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 79:299–318, 2008.
[53] Liu B, Yu H, Zhou X, Liu H, Combining 3d joints moving trend and geometry property for human action recognition, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 1–6, 2016.
[54] Chang CC, Lin CJ, Libsvm: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2:1–27, 2011.
[55] Li W, Zhang Z, Liu Z, Action recognition based on a bag of 3d points, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 9–14, 2010.
[56] Ju Z, Ji X, Li J, Liu H, An integrative framework of human hand gesture segmentation for human-robot interaction, IEEE Systems Journal 99:1–11, 2015.

[57] Wang Y, Yu H, Stevens B, Liu H, Dynamic facial expression recognition using local patch and lbp-top, Proceedings of the 8th International Conference on Human System Interactions, 362–367, 2015.
[58] Ojala T, Pietikainen M, Maenpaa T, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24:971–987, 2002.
[59] Zhao G, Pietikainen M, Dynamic texture recognition using local binary patterns with an application to facial expression, IEEE Transactions on Pattern Analysis and Machine Intelligence 29:915–928, 2007.
[60] Agarwal L, Lakhwani K, Optimization of frame rate in real time object detection and tracking, International Journal of Scientific & Technology Research, 2:132–134, 2013.
[61] Zhou X, Yu H, Liu H, Li YF, Tracking multiple video targets with an improved GM-PHD tracker, Sensors 15(12):30240–30260, 2015.
[62] Zhou X, Li YF, He B, Entropy distribution and coverage rate-based birth intensity estimation in GM-PHD filter for multi-target visual tracking, Signal Processing 94:650–660, 2014.
[63] Zhou X, Li YF, He B, Bai T, GM-PHD-based multi-target visual tracking using entropy distribution and game theory, IEEE Transactions on Industrial Informatics 10:1064–1076, 2014.
[64] Zhang S, Yu H, Dong J, Wang T, Qi L, Liu H, Combining kinect and pnp for camera pose estimation, Proceedings of 8th International Conference on Human System Interactions, 357–361, 2015.
[65] Kumatani K, Arakawa T, Yamamoto K, McDonough J, Raj B, Singh R and Tashev I, Microphone array processing for distant speech recognition: Towards real-world deployment. In IEEE Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1–10, 2012.
[66] Tashev, I, Recent advances in human-machine interfaces for gaming and entertainment. International journal of information technologies and security 3(3):69–76, 2011.
[67] Kinnunen T, Li H, An overview of text-independent speaker recognition: From features to supervectors, Speech communication 52:12–40, 2010.
[68] Drejing K, Thill S, Hemeren P, Engagement: A traceable motivational concept in human-robot interaction, Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 956–961, 2015.
[69] Sloman A., Beyond shallow models of emotion, Cognitive Processing 2(1):177–198, 2001.
[70] Norman DA, Ortony A and Russell DM, Affect and machine design: Lessons for the development of autonomous machines, IBM Systems Journal 42(1):38–44, 2003.
[71] Belpaeme T, Baxter P, Read R, Wood R, Cuayáhuitl H, Kiefer B, Racioppa S, Kruijff-Korbayová I, Athanasopoulos G, Enescu V and Looije R, Multimodal child-robot interaction: Building social bonds. Journal of Human-Robot Interaction, 1(2):33–53, 2012.
[72] Gómez Esteban P, Cao HL, De Beir A, Van de Perre G, Lefeber D, Vanderborght B, A multilayer reactive system for robots interacting with children with autism, ArXiv preprint arXiv:1606.03875, 2016.
[73] Fujiwara K, Kanehiro F, Kajita S, Kaneko K, Yokoi K, Hirukawa H, Ukemi: falling motion control to minimize damage to biped humanoid robot, Proceeding of International Conference on Intelligent Robots and Systems, 2521–2526, 2002.
[74] Yun SK, Goswami A, Hardware experiments of humanoid robot safe fall using Aldebaran Nao, Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 71–78, 2012.
[75] Ford C, Bugmann G, Culverhouse P, Modeling the human blink: A computational model for use within human–robot interaction, International Journal of Humanoid Robotics 10(1), 2013.
[76] Ferreira JF, and Dias J, Attentional Mechanisms for Socially Interactive Robots–A Survey, IEEE Transactions on Autonomous Mental Development 6(2):110–125, 2014.
[77] Zaraki A, Mazzei D, Giuliani M, De Rossi D, Designing and evaluating a social gaze-control system for a humanoid robot, IEEE Transactions on Human-Machine Systems 44(2):157–168, 2014.
[78] Lanillos P, Ferreira JF, and Dias J, Designing an artificial attention system for social robots, In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 4171–4178, 2015.
[79] Sutton RS and Barto AG, Reinforcement learning: An introduction. 1(1), 1998, Cambridge: MIT press.
[80] LeCun Y, Bengio Y, and Hinton G, Deep learning. Nature, 521(7553):436–444, 2015.
[81] Amershi S, Cakmak M, Knox WB, and Kulesza T, Power to the people: The role of humans in interactive machine learning. AI Magazine, 35(4):105–120, 2014.
[82] Fails JA and Olsen DR, Interactive machine learning, Proceedings of the 8th international conference on Intelligent user interfaces. ACM, 2003.
[83] Argall BD, Chernova S, Veloso M, and Browning B, A survey of robot learning from demonstration. Robotics and autonomous systems, 57(5):469–483, 2009.
[84] Billard A, Calinon S, Dillmann R, and Schaal S, Robot programming by demonstration. In Springer handbook of robotics, 1371–1394, 2008.
[85] Senft E, Baxter P, Kennedy J, Belpaeme T, Sparc: Supervised progressively autonomous robot competencies, Proceedings of International Conference on Social Robotics, 603–612, 2015.
[86] Senft E, Baxter P, Belpaeme T, Human-guided learning of social action selection for robot-assisted therapy, 4th Workshop on Machine Learning for Interactive Systems, 2015.
[87] Senft E, Lemaignan S, Baxter P and Belpaeme, T, SPARC: an efficient way to combine reinforcement learning and supervised autonomy, Future of Interactive Learning Machines workshop, 2016.
[88] Thomaz AL, and Breazeal C, Teachable robots: Understanding human teaching behavior to build more effective robot learners, Artificial Intelligence 172(6):716–737, 2008.
[89] Sloman A, Varieties of Meta-cognition in Natural and Artificial Systems, In AAAI Workshop on Metareasoning, 8:12-–20, 2011.
[90] Coeckelbergh M, Are emotional robots deceptive?, IEEE Transactions on Affective Computing, 3(4):388–393, 2012.
[91] Richardson K, An Anthropology of Robots and AI: Annihilation Anxiety and Machines, Routledge, 2015.
[92] Anderson M, Anderson SL, Machine ethics, Cambridge University Press, 2011.
[93] Eurobarometer, Public attitudes towards robots, 2012, http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf
[94] Itoh K, Miwa H, Matsumoto M, Zecca M, Takanobu H, Roccella S, Carrozza M, Dario P, Takanishi A, Various emotional expressions with emotion expression humanoid robot we-4rii, Proceedings of IEEE technical exhibition based conference on robotics and automation, 35–36, 2004.

[95] Sugiyama O, Kanda T, Imai M, Ishiguro H, Hagita N, Natural deictic communication with humanoid robots, Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 1441–1448, 2007.

[96] Ido J, Matsumoto Y, Ogasawara T, Nisimura R, Humanoid with interaction ability using vision and speech information, Proceedings of IEEE/RSJ International conference on Intelligent Robots and Systems, 1316–1321, 2006.

[97] Zecca M, Mizoguchi Y, Endo K, Iida F, Kawabata Y, Endo N, Itoh K, Takanishi A, Whole body emotion expressions for kobian humanoid robot: preliminary experiments with different emotional patterns, Proceedings of the 18th IEEE International symposium on robot and human interactive communication, 381–386, 2009.

[98] Ekman P, Friesen W, Facial Action Coding System, Consulting Psychologists Press, 1978.

[99] Saldien J, Goris K, Vanderborght B, Vanderfaeillie J, Lefeber D, Expressing emotions with the social robot probo, International Journal of Social Robotics 2(4):377–389, 2010.

[100] De Beir A, Cao HL, Gómez Esteban P, Van De Perre G, Vanderborght B, Enhancing Nao Expression of Emotions Using Pluggable Eyebrows, International Journal of Social Robotics, 1-9, 2015.

[101] Atkinson AP, Dittrich WH, Gemmell AJ, Young AW, et al, Emotion perception from dynamic and static body expressions in point-light and full-light displays, Perception-London 33(6):717–746, 2004.

[102] Van de Perre G, Van Damme M, Lefeber D, Vanderborght B, Development of a generic method to generate upper-body emotional expressions for different social robots, Advanced Robotics 29(9):597–609, 2015.

[103] RocketBox (2016) http://www.rocketbox-libraries.com

[104] Hirai K, Hirose M, Haikawa Y, Takenaka T, The development of honda humanoid robot, Proceedings of the 1998 IEEE International Conference on Robotics and Automation, 2:1321–1326, 1998.

[105] Van de Perre G, De Beir A, Cao HL, Esteban PG, Lefeber D, and Vanderborght B, Reaching and pointing gestures calculated by a generic gesture system for social robots. Robotics and Autonomous Systems, 83:32–43, 2016.

[106] Yussof H, Salleh MH, Miskam MA, Shamsuddin S, Omar AR, Asknao apps targeting at social skills development for children with autism, Proceedings of IEEE International Conference on Automation Science and Engineering (CASE), 973–978, 2015.