Research Article **Open Access**

Kim Baraka*, Francisco S. Melo, and Manuela Veloso

Interactive robots with model-based 'autism-like' behaviors

Assessing validity and potential benefits

https://doi.org/10.1515/pjbr-2019-0011
Received June 30, 2018; accepted January 26, 2019

Abstract: Due to their predictability, controllability, and simple social abilities, robots are starting to be used in diverse ways to assist individuals with Autism Spectrum Disorder (ASD). In this work, we investigate an alternative and novel research direction for using robots in relation to ASD, through programming a humanoid robot to exhibit behaviors similar to those observed in children with ASD. We designed 16 'autism-like' behaviors of different severities on a NAO robot, based on ADOS-2, the gold standard for ASD diagnosis. Our behaviors span four dimensions, verbal and non-verbal, and correspond to a spectrum of typical ASD responses to 3 different stimulus families inspired by standard diagnostic tasks. We integrated these behaviors in an autonomous agent running on the robot, with which humans can continuously interact through predefined stimuli. Through user-controllable features, we allow for 256 unique customizations of the robot's behavioral profile. We evaluated the validity of our interactive robot both in video-based and 'in situ' studies with 3 therapists. We also present subjective evaluations on the potential benefits of such robots to complement existing therapist training, as well as to enable novel tasks for ASD therapy.

Keywords: socially interactive robots, autism spectrum disorder, autism diagnostic observation schedule, autism-like robot behaviors, training simulation

***Corresponding Author: Kim Baraka:** Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA / INESC-ID/Instituto Superior Técnico, Universidade de Lisboa, 2744-016 Porto Salvo, Portugal; E-mail: kbaraka@andrew.cmu.edu
Francisco S. Melo: INESC-ID/Instituto Superior Técnico, Universidade de Lisboa, 2744-016 Porto Salvo, Portugal; E-mail: fmelo@inesc-id.pt
Manuela Veloso: School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: mmv@cs.cmu.edu

1 Introduction

1.1 Background and scope

Autism Spectrum Disorder (ASD) is a developmental condition that affects individuals' communication and social abilities, as well as possibly motor and cognitive skills. The behavioral profiles of individuals with ASD span an extremely diverse spectrum, resulting in a large variability and individuality of resulting behavioral profiles [1]. Even though ASD is being studied from very different perspectives, including developmental, neurophysiological [2], and genetic ones [3, 4], its diagnosis primarily relies on *behavioral* observation in controlled settings.

As a result, available diagnostic tools for ASD used by therapists provide us with behavioral models for ASD. More specifically, these tools link a taxonomy of typically observed behaviors to values on a set of features that have been identified to be relevant to characterizing the condition in its diverse forms. In particular, the Autism Diagnosis Observation Schedule (ADOS-2) [5] is a state-of-the-art tool for diagnosis through interaction and observation of a child's behaviors in a semi-controlled environment. The therapists go through a series of 10 tasks with the child using standardized objects and procedures, then code the behaviors they observed throughout the session in the form of discrete values on a set of features spanning several behavioral dimensions. A typical ADOS-2 session takes 40–60 minutes to administer. Different modules are available depending on the child's age or language ability. In this work, we focus on Module 2, suitable for children with phrase speech abilities, which provides us with a richer set of behaviors as compared to the other existing modules. Figure 1 shows a sample ADOS-2 feature and task.

This work builds and expands on our previous research [6], whose goal is to apply the ADOS-2 model to control the behaviors of a humanoid NAO robot, enabling it to exhibit behaviors similar to those of children with varying severities of ASD. Autonomous robots and agents

Feature 'Overall Level of Non-Echoed Spoken Language'	*Task 'Response to Name': Hierarchy of presses*
0 = Three or more word utterances; some grammatical markings.	(1) Call the child's name up to 4 times
1 = Two or three word utterances; no grammatical markings.	(2) Ask the caregiver to call the child's name up to 2 times
2 = Occasional phrases; mostly single words.	(3) Ask the caregiver to make a familiar noise or sound up to 2 times
3 = Echoed speech only	(4) Touch the child to catch his/her attention

Figure 1: A simplified example of a feature (left) and task (right) from Module 2 of the ADOS-2. Coding schemes on different features possess variable levels of subjectivity (the one shown being highly objective). Hierarchies of presses are traversed sequentially until the desired child behavior is observed. (Figure adapted from [5], ©WPS.)

are commonly and informally described as 'autistic' [7, 8] in some of the literature (referring to their lack of social intelligence). However, the belief that behaviors in individuals with ASD are less complex versions of behaviors in typically developing individuals isn't always true. In fact, ASD may introduce interesting and rich subtleties, idiosyncrasies, and proactive behaviors not seen in typically developing individuals. On a more global note, calling robots or agents 'autistic' may be dangerous because it reinforces erroneous assumptions on how the autistic mind functions [9]. In this work we use the term 'autism-like' to refer merely to the robot behaviors themselves. These visually resemble human behaviors typically observed in children with ASD. They are modeled purely at the behavioral level, not considering any simulation of lower-level cognitive processes that individuals with ASD may possess.

Because of the great diversity of behavioral profiles possible on the robot, we introduce a way for the user to customize the model parameters in order to allow for different severities of ASD along 4 behavioral features, both verbal and non-verbal. We start by designing 'autism-like' behaviors of varying severities along those features, on a NAO robot, based on the ADOS-2 model. Those behaviors correspond to a range of possible responses to three different stimulus families, inspired by the ADOS-2 tasks. In a second step, we integrate those behaviors in an autonomous agent running on the robot. The result is a customizable interactive robot with 'autism-like' behaviors, capable of continuous interaction with a human in response to a set of predefined stimuli. Through the specification of user-controllable features, the robot can be customized in one of 256 unique ways. We evaluated the validity of the resulting interactions both in video-based and 'in situ' studies with certified autism therapists.

To summarize, the main contributions of this work are the following:

- A method for controlling the behavioral responses of a NAO humanoid robot with 'autism-like' behavioral characteristics, based on the ADOS-2 diagnostic tool,
- A set of 16 robot behaviors, spanning different ADOS-2 features and different severities, which include speech, gaze and gestures, in response to verbal, sound, and touch stimuli,
- An architecture for integrating these behaviors in a costumizable interactive agent running on the robot, and
- An evaluation, through video-based and 'in situ' studies, of the validity and potential benefits of our approach for complementing therapist training, as well as helping with novel ASD tasks.

In the next subsection, we discuss in greater detail the relevance and potential impact of these different contributions

1.2 Why robots with 'autism-like' behaviors?

Most existing research on the use of robots for autism has focused on assisting individuals with ASD directly, mainly in therapeutic settings [10, 11]. The rationale is that the predictability, controllability, and simplicity of robots' social skills can benefit such a population by engaging them into simplified social interactions that would hopefully generalize to real-world interactions. In this work, we propose an alternative way of using robots in relation to ASD, through enabling robots themselves to emulate typical ASD behaviors. We foresee several real-world applications that motivate the use of robots with 'autism-like' behaviors, including complementing therapist training and enabling new types of autism therapy tasks. We will now motivate each of those applications separately.

1.2.1 Therapist training

Current therapist training for ASD diagnosis procedures[2] heavily relies on videos and theoretical material, as well as observing a real diagnosis session run by a trained expert. Even though it exposes the therapists in training to a wide range of examples of behaviors and stresses on the rigorousness of the coding schemes and task procedures,

2 https://www.wpspublish.com/store/c/343

it largely ignores the interactive and embodied component required for a successful administration of the tool. In fact, this component represents a crucial part of the administration process. Given that therapists are expected to follow very specific sets of instructions (e.g., Figure 1), while paying attention to behaviors, taking notes, and possibly adapting the order of tasks in real-time, a poor mastering of these interactive skills may result in mistakes in task administration as well as feature coding. A lowered reliability, especially in the coding of some features with already low agreement scores [5] defeats the purpose of using a standardized tool in the first place. On the other hand, utilizing robots capable of exhibiting 'autism-like' behaviors way may possibly help to complement the existing training by providing an interactive simulation environment for therapists to train on before moving on to real scenarios.

The paradigm of using simulated environments or interactions for expert training [12, 13] has already been applied to a wide range of fields, including aviation [14], medicine and healthcare [15, 16], the military [17], emergency response [18], and education [19, 20], among others, showing improvement in the performance of trainees in most cases. Simulated environments have also been applied to social settings and interactions [21–23], as well as procedural tasks [24]. The large majority of these solutions rely on computer simulations and virtual/mixed reality, while very little work has been done on the introduction of embodied agents in these simulated environments. The only work we found that used some sort of physical feedback or embodied communication was for welding [25] and surgical procedures [15]. To the best of our knowledge, the use of social robots in the context of professional therapist training has not been yet investigated.

An interactive robot capable of emulating 'autism-like' behaviors has the following advantages:

- *Interactivity*: Unlike existing therapist training, a robot is capable of emulating, to a limited extent, the structured interactions of an ADOS-2 session.
- *Customizability*: In the real world, the experience therapists gather depends on the patients they receive, which is not controllable. The customizable aspect of our robot allows to generate arbitrary behavioral profiles, greatly increasing the number and diversity of feature combinations the therapists can be exposed to.
- *Repeatability*: Real-life interactions happen only once, and if we attempt to repeat them, there will always be some inevitable differences. Even though videos showing behaviors or interactions may be repeated to be better studied, the use of an interactive robot allows the interaction itself to be repeated in a controlled way, and allows for reiterating previous interactions in the event of procedural errors, or lack of observational attention.

Furthermore, research on human perception has shown that people tend to assign human-like traits to technological artifacts, including robots, perceiving them as social beings [26]. This aspect of our cognition motivates the use of humanoid robots that do not necessarily have to reproduce the physical appearance or size of a child with high fidelity. In fact, the humanoid NAO robot used in this work is much smaller than a human being, but possesses basic features that make it expressive and able to exhibit engaging social behaviors.

1.2.2 Autism therapy

While therapist training is the main focus of the studies presented in this work, we believe that ASD therapy may also benefit from having a robot capable of exhibiting 'autism-like' behaviors. Specifically, these robots may unlock new possibilities in robot-assisted therapy tasks involving imitation, as well as learning-by-teaching scenarios, as discussed next.

Imitation tasks hold a special place in ASD therapy [27], because imitation ability is often impaired in children with ASD [28]. As a result, we believe that an autonomous, customizable, and adaptive robot that is able to match its behavior to that of the child, demonstrate a desirable behavior for the child to imitate, or, in the context of long-term interaction, evolve towards less severe behaviors along with the child, may hold promise in the context of ASD therapy.

On the other hand, such a robot may be used in the context of learning-by-teaching scenarios [29], where a child refines his/her own skills through teaching the robot already acquired skills. For example, the robot could be programmed to have slightly lower skills than the child (i.e., higher severity on a given a feature), in which case the child teaches the robot to incorporate modalities that the robot doesn't have. For instance, if the child knows how to make good use of pointing, he/she could teach a robot that uses only eye gaze to also include pointing in its behavior. However, it is to be noted that a learning-by-teaching approach might be challenging with some children on the higher end of the autism spectrum, and would require empirical investigation.

The motivating thoughts of this section do not provide empirical evidence for the usefulness of these applications in the associated contexts, but rather are meant to

provide a basis for the conception of future robot-assisted scenarios in the autism domain.

In the existing literature, customizable social robots have been mainly developed to account for user differences and preferences [31–33], although personalization for individuals with ASD [34] has not yet been thoroughly investigated. In relation to emulation of typical ASD behaviors by robots, some work has been done on real-time motion imitation of children with ASD [35]. Additionally, some research looked at using robots as a platform to test theories related low-level cognitive and sensorimotor processes related to ASD, in relation to specific aspects of behavior such as sensory integration and movement [36] or joint attention [37]. However, to the best of our knowledge, enabling humanoid robots to exhibit 'autism-like' behaviors along severity scales, based on standardized behavioral models such as the ADOS-2, has never been looked at before.

The rest of this article is organized as follows. Section 2 describes our approach to designing customizable and autonomous robots with 'autism-like' behaviors, as well as the video-based and 'in situ' evaluation studies we conducted. Section 3 presents and discusses the results of our two studies, and section 4 concludes and presents some future work directions.

2 Methods

We first describe our methods for designing an interactive, autonomous and customizable robot with 'autism-like' behaviors. We then present two studies (video-based and 'in situ') to validate our behavior design, evaluate interactivity, and assess potential real-world benefits of our solution.

2.1 Designing a customizable interactive robot with 'autism-like' behaviors

Based on the ADOS-2 model, we designed 16 behaviors on a NAO robot and integrated them into an autonomous agent architecture that can be customized according user-specified feature values. The robot is able to automatically detect some interaction parameters, such as verbal and non-verbal stimuli, as well as sound location, to allow for more natural interactions.

2.1.1 Robot behaviors based on the ADOS-2 model

We selected 4 **features** from the ADOS-2 Module 2 to inform our design of robotic behaviors that emulate those of children with varying ASD severities. The features are: 'Response to name', 'Response to joint attention', 'Overall level of non-echoed speech', and 'Pointing'. As with a child, those features can characterize the responses of our robot to different stimuli. We consider three hierarchical **stimulus families**, namely: calling attention by calling the name (N), calling attention towards an object (JA), and asking for snack preference (S), each of which contains a set of stimuli with the same intention or purpose. Those stimuli, inspired by the hierarchical 'presses' of the ADOS-2 tasks, are summarized in Figure 2.

The 4 features we selected can each take on discrete values between 0 and 3, corresponding to ASD **severities** along the corresponding feature (in other words, higher values are associated with more autistically severe behaviors). The ADOS-2 manual provides a detailed description of the sets of behaviors that correspond to each feature severity, and we utilize those descriptions to design 16 robot **behaviors**, consisting of, for each separate feature severity, one selected behavior that was easily reproducible on a robot. A robot behavior consists of an animation of the robot's joints as well as possibly speech, and is triggered by a subset of the stimuli we defined. Some of our behaviors are parametrized (e.g., gaze behavior takes as a parameter a 3D location to look at). Table 1 presents a summary of our designed behaviors, in response to the three stimulus families N, JA, and S. In the presence of more than one relevant feature for a stimulus family (e.g., S), behaviors are **blended**, meaning they are run simultaneously.

2.1.2 Integration into an autonomous agent architecture

Our designed behaviors were integrated as part of an autonomous agent capable of having continuous interactions with one or more humans, according to the predefined stimuli it recognizes. More importantly, the agent can be customized by specifying an arbitrary severity for each feature, resulting in 256 unique customizations. The architecture of the autonomous agent, including a perception module with speech recognition, touch recognition, and sound localization to modulate the robot behaviors, is summarized in Figure 3. We implemented this architec-

Figure 2: Example of stimulus families considered in this work, inspired by the 'presses' of the ADOS-2 tool. Shown pictures are for the 'Calling name' family. For the 'Calling for joint attention' family, the stimuli are: Verbal stimulus: 'Look!' - Verbal stimulus: 'Look at THAT!' - Activating the object. For the 'Asking for snack preference' family, the only stimulus is the verbal stimulus: 'Which snack do you like?'.

Table 1: Summary of the designed 'autism-like' robot behaviors of varying severities.

Stimulus family	Relevant feature(s)	Responses			
		Severity 0	Severity 1	Severity 2	Severity 3
Calling name (N)	Response to name (rN)	Looks at human within second name calling attempt with coordinated utterance "Yes?" (rN0)	Same as rN0 but only responds to 'familiar' human while ignoring 'non-familiar' one (rN1)	Looks in general direction (without eye contact or utterances) of 'familiar' human only while ignoring 'non-familiar' one (rN2)	Only responds to touch on head by exhibiting succession of random gaze shifts; ignores all other stimuli in N (rN3)
Calling for joint attention (JA)	Response to joint attention (rJA)	Immediately looks at object, then human, then back at object (rJA0)	Ignores first stimulus; looks at object only at second stimulus "Look at THAT!" (rJA1)	Ignores first two stimuli; only looks at object when activated and emitting sound (rJA2)	Same as rJA2 but with slight gaze shift towards object without actually looking at object (rJA3)
Asking for snack preference (S)	Overall level of non-echoed speech (rlS)	Says: "I like this snack of all the snacks in the world." (rlS0)	Says: "This one." (rlS1)	Says: "This." (rlS2)	Echoes: "Snack... Snack... Snack... Like... Like..." (rlS3)
	Pointing (rpS)	Clearly points at one of the snacks with coordinated eye gaze (rpS0)	Clearly points at one of the snacks with slight gaze shift not in direction of pointing (rpS1)	Looks at one of the snack but without pointing (rpS2)	Slightly shifts gaze downwards with no pointing (rpS3)

ture on the NAO robot using the NAOqi Python API through the Choregraphe suite[3].

Because of perceptual limitations of the robot, some parameters needed to be hardcoded or estimated simplistically, while others are easier to detect completely autonomously. Below are some more details on the parameters automatically estimated vs. hardcoded, for each behavior, following the abbreviations of Table 1:

- rN0 through rN3: Voice location is estimated using NAO's microphone array and used to modulate the eye gaze of the robot. The 'familiar' and 'unfamiliar' humans are distinguished simplistically, based on the location of the voice. It is assumed that the 'familiar' person would always be on one side of the robot (e.g., left) and the 'unfamiliar' always on the other (e.g., right). The touch sensor on NAO's head is used for rN3.
- rJA0 through rJA3: Because of the robot's perceptual limitations, the location of the object used for calling joint attention is hardcoded in rJA0 and rJA1. For rJA2 and rJA3, it is estimated using sound localization, since the activated object emits a sound. For motion stability purposes (robot loosing balance at times), the location of the human in the joint attention task is hardcoded.

3 Code available at https://github.com/kobotics/autistic_nao

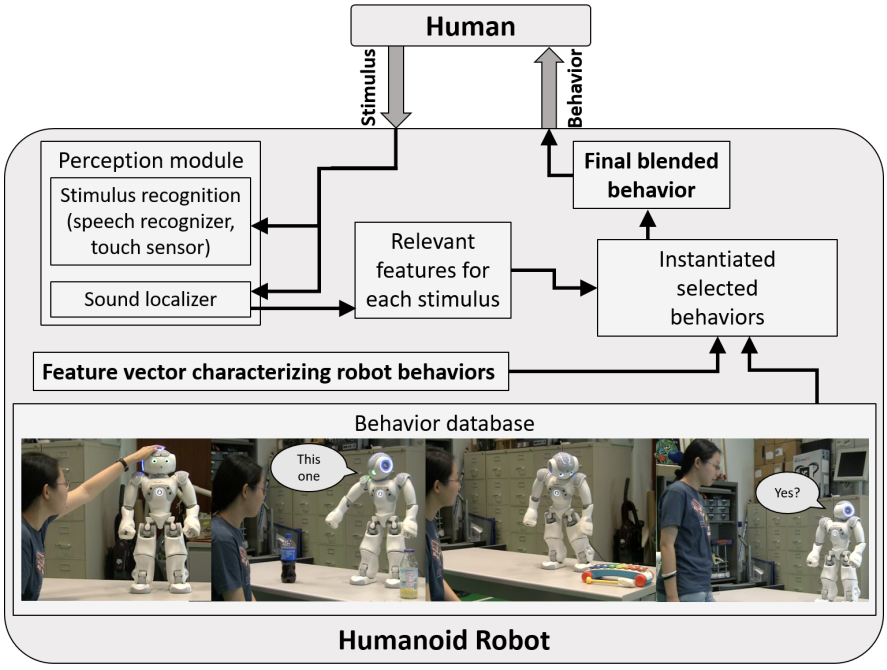

Figure 3: Architecture of our customizable autonomous agent; stimuli are recognized and trigger different behaviors, according to the customizable feature vector characterizing the robot.

- rpS0 through rpS3: Two snacks were put on the table, whose positions are hardcoded. The preferred snack position is used to parametrize the eye gaze and pointing directions of the robot.
- rlS0 through rlS3: These behaviors consist of speech only, and are not parametrized.

Note that, for all behaviors, the speech recognizer is used to detect all verbal stimuli, which triggers the corresponding responses, when applicable. When idle, the robot is animated through a subtle 'Breathing' behavior in which the robot slightly shifts its weight from one foot to the other. A video showing sample human interactions with our autonomous NAO robot in 'low severity' vs. 'high severity' modes is available for online viewing[4].

2.2 Evaluating the designed behaviors (video-based study)

In order to evaluate the validity of our designed interactive behaviors with respect to the formalism of the ADOS-2, we ran a first video-based evaluation study with trained ASD therapists. The aim of the study was to investigate: (1) whether the therapists would assign to the features characterizing the designed behaviors the same values as the ones on which their design was based, and (2) whether the therapists would agree with each other in their evaluation, and how this agreement would differ across the different behavioral responses of the robot.

2.2.1 Survey structure

The study consisted of a video-based survey showing short videos[5] of the isolated designed behaviors in the context of an interaction with a human (or two for the behaviors requiring more than one person). Based on what they saw in the video, the participants provided a severity value between 0 and 3 on the relevant feature(s) of each video, according to the description for each severity level in the ADOS-2 manual. Detailed instructions were given in relation to feature coding, background on robot's capabilities, and simplifying assumptions. In particular, the participants were instructed to 'diagnose' the robot the same way they usually do it with children, by coding the feature value they thought best characterized the response they observed in the video. They had the possibil-

4 Video available at https://bit.ly/2tE2nOD

5 Survey videos available at https://bit.ly/2MqRdDK

ity to watch the video as many times as needed. Also, they were instructed to use information from the current video only, and after the first stimulus was started (even though some of the features usually require several samples to form a good judgment). Finally, they were asked to ignore any expression unrelated to motion or speech, including non-verbal cues acknowledging the detection of speech, namely beeps and color changes of the NAO's eyes. These cues, part of the default behavior of the speech recognizer, were kept in our interaction because they were designed to facilitate speech synchronization and the debugging of the state of the robot in case of a recognition failure.

The videos were organized into three tasks (N, JA, and S), corresponding to the stimulus families discussed in section 2.1. Because behaviors were blended in task S, and to avoid overwhelming the participants with a very large number of videos, we chose to set the feature values to be identical, in all videos for that task, for both language and pointing features (i.e., rlS0,rpS0 - rlS1,rpS1 - ...). The total number of videos was hence 12, four for each of the three tasks. The order of the three tasks in the survey was randomized, as well as the order of the videos within each task. When applicable, the progression of stimuli was performed in the hierarchical order used in the ADOS-2 presses until a response was seen on the robot. The survey also included snapshots of the corresponding ADOS-2 manual to help the trained experts code the severity on each feature based on their observations.

2.2.2 Methodology

We first ran a small pilot with one trained therapist to get an idea of the expected results, as well as gather feedback on the clarity of the survey, and potential points for improvement. When the survey was finalized, we gathered the online responses of three other therapists from the Child Development Center at the Hospital Garcia de Orta in Almada, Portugal. The therapists who participated in this study are all women who received a training in administering ADOS-2. Informed consent and permission to use media was obtained at the beginning of the survey.

2.3 Assessing therapist-robot interaction and potential benefits ('in situ' study)

The aim of this second study was to test our robot under different configurations in a real interactive setting with autism therapists, as well as assessing the potential benefits of this interactive robotic tool. This study therefore relied on, first, coding of robot behaviors according to the ADOS-2 specifications and, second, answering a questionnaire we devised to assess the potential benefits of our robot in real-world applications. This study was performed with the same three participants from the video-based study, 11 months later.

2.3.1 Methodology

In the main part of this study, the participants interacted with the robot through the set of stimulus families we defined, observing and subsequently coding the robot's responses according to the ADOS-2 specifications, as was done in the previous study (described in section 2.2). The robot configurations were similar as well (matching severities on language and pointing features), and we exposed the participants to the same 12 robot responses. However, there were some differences as compared to the video-based study:

- In the video-based study, we consecutively showed different robot responses for the same stimuli to allow for better comparison of behaviors, as the focus was solely on validating the behaviors themselves. In this study, we are interested in a more naturalistic evaluation of the interaction as a whole that goes beyond isolated behaviors. As a result, we had the participants go through each task once, then repeat the process, with 4 different robot customizations randomly permuted while ensuring that each robot behavior appeared once. This way, participants could get a sense of an entire interaction with 4 'different' robots that they would have to diagnose, similar to a real ADOS-2 session.
- As participants were allowed to replay the videos in the previous study, in this study, they were allowed to repeat the task as many times as needed for coding the behaviors.
- Within the constraints imposed by our robot, we tried to replicate as much as possible the physical setting that the therapists are used to. For example, we used objects from the ADOS-2 kit, such as one of the activatable toys from ADOS-2, one savory and one sweet snack, which differed slightly from the ones used the videos.

In addition to coding behaviors, we also asked participants to provide answers to a questionnaire, separated into two parts. The aim of this questionnaire was to compare the ratings of existing training solutions with our proposed solution, as well as to evaluate the potential bene-

fits of robots with 'autism-like' behaviors in our foreseen applications.

The first part of the questionnaire, presented before the interaction with the robot, first gathered background information about the participant's diagnostic training. It then asked the participants to assess that training along three dimensions, namely:

1. *Behavior accuracy*, i.e., comparing behaviors encountered in training vs. encountered in real sessions;
2. *Interactivity*, i.e., to what extent it involves an interaction in real or virtual scenarios;
3. *Diversity of behavioral profiles*, i.e., diversity of combinations of severities on the ADOS-2 features.

Finally, it asked for how much they believed robots with 'autism-like' behaviors could benefit the our foreseen applications, namely: complementing existing ADOS-2 *training of therapists*, and enabling new types of scenarios for *autism therapy* (e.g., imitation tasks). In addition, we included as a third potential application *educating and sensitizing* the general population about the behavioral differences in children with ASD (e.g., classrooms, museums, workplace,...).

The second part of the questionnaire, presented after the interaction with the robot, repeated the same questions as the first part, but this time assessing specifically our robotic tool. The questionnaire structure is summarized in Figure 4. Apart from the 'Training background' section, which was multiple choice, all responses were in the form of a 5-point Likert scale. The questionnaire was in the participants' native (Portuguese) language.

- **Training background**
  - Type of training received
  - Experience with general tool
  - Experience with module used in this study
  - Content of training received
- **Training assessment**
  - Behaviors encountered in training vs. behaviors encountered in real session?
  - Interactivity?
  - Diversity of behaviors profiles?
  - Other?
- **Envisioned benefits of robots with 'autism-like' behaviors (general)**
  - Therapist training?
  - Autism therapy?
  - Education and sensitization
  - Other?

[Interaction with the robot and coding; briefly show the interface and explain how it works]

- **Evaluation of the current tool**
  - Behaviors encountered in training vs. behaviors encountered in real session?
  - Interactivity?
  - Diversity of behaviors profiles?
  - Other?
- **Envisioned benefits of our robotic tool**
  - Therapist training?
  - Autism therapy?
  - Education and sensitization
  - Other?

Figure 4: Questionnaire structure. An English version of the full questionnaire is available at https://bit.ly/2KhqnRV.

2.3.2 Procedure

After signing an informed consent, the participants filled the first part of the questionnaire. The examiner then took them into the robot experiment room and provided them with instructions on the tasks they were going to perform on the robot, as well as the structure of the rest of the study. Sheets with all needed information were made available to them, including the list of valid stimuli for each task, relevant snapshots from the ADOS-2, and space to use for coding. In addition to notes mentioned in the video-based study, the examiner also stressed that it was important that they spoke clearly and loudly, and that the robot only responds to voice and touch but not visual cues such as gaze direction of pointing. As in the video-based survey, the examiner reminded the participants to only consider in their coding the robot's behavior after the first stimulus of a given task was started. We also asked them to ignore any robot expressions that are unrelated to motion or speech.

Once any doubts they had was clarified, they 'diagnosed' the robot with the first customization going through the three tasks sequentially N - JA - S, observing the robot's responses and reporting their codes on the sheet. Once the three tasks were over, the examiner announced that he was going to reprogram the robot, and asked the participant to treat it as a 'new robot'. The process was repeated until all 4 pre-randomized robot customizations were shown. Figure 5 shows some snapshots of these interactions.

Because of technical limitations of the robot, there were moments where the examiner had to briefly intervene, saying things like "the robot didn't understand what you said, please repeat". The examiner, although present in case of doubt from the participant's part, tried to be as non-invasive as possible to maintain the naturalness of the interaction.

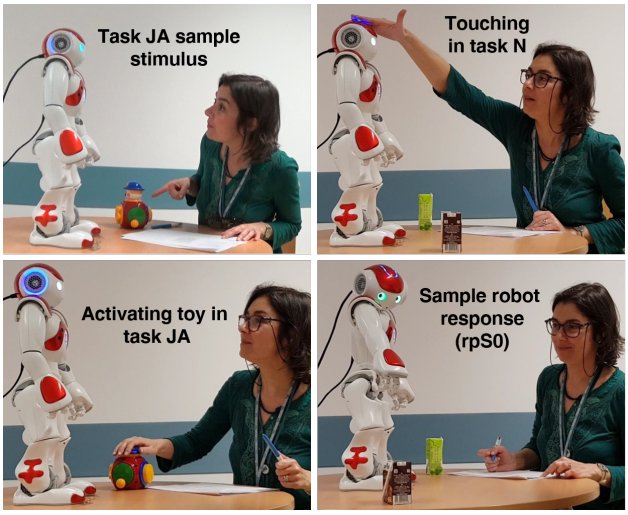

Figure 5: Autism therapist interacting with the robot and coding its responses according to the ADOS-2 specifications.

3 Results and discussion

We first analyzed, across the two studies, the accuracy of responses and the agreement between the participants, summarized in Tables 2 and 3. In our accuracy analysis, we only discriminated between correctly and incorrectly classified responses (with respect to the expected response). In the agreement analysis, we also treated the variables as ordinal in one of our metrics, and compared some of the results with reference values from real ADOS-2 scenarios. We additionally investigated order effects in the responses. Finally, we analyzed the results of the questionnaire and compiled additional qualitative observations.

3.1 Accuracy results

The participants achieved an overall accuracy of 76.04% across the two studies. The accuracy was considerably higher for the video-based study (83.33%), as compared to the 'in situ' study (68.75%), with close to statistical significance using a McNemar's mid-p test on the overall binary categorical data ($p = 0.057$), and actual statistical significance only for expert 1 ($p = 0.031$). The same test showed no statistical significant difference in accuracy between all pairs of raters ($p \geq 0.125$ for all pairs), regardless of the type of interaction (video/real). There also seems to be a relationship between the level of experience and the accuracy of the participants. The accuracy results are summarized in Table 2 and the left side of Table 3.

Looking at individual features, the joint attention feature (rJA) had the highest accuracy (91.67%), and the pointing feature (rpS) the lowest (54.17%), which was expected because the latter was the most complex one to code, and was paired with speech behavior within the same task, possibly resulting in some interaction effects. However, the hypothesis that blending behaviors from supposedly independent features may involve interaction effects in the coding of individual features needs to be investigated more carefully with a larger sample. Moreover, we expected the language feature (rlS) to have the highest accuracy as it was the least subjective feature to code, which wasn't the case. We hypothesize that it may have also been subject to the interaction effect discussed above, as well as the fact that it was the only feature that differed considerably between Module 2 and Module 1 of the ADOS-2, the latter being the one that the participants were most used to in their professional practice.

To understand better the sources of misclassifications, we report more granular accuracy results in Figure 6, for each robot behavior. We can see that 11 out of 16 behaviors had an overall accuracy superior to 80%. Note that the joint attention feature (rJA) showed relatively high accuracy for all behaviors. On the other hand, the two behaviors which had the lowest accuracies were rpS1 and rpS2. In some cases, it seemed that the participants thought it would be appropriate to code the gaze behavior as part of the pointing behavior, which shouldn't have happened given that eye gaze is typically coded in a separate feature (not included in these studies). In other cases, on the contrary, participants seemed to have completely denied the importance of gaze for the pointing feature, which is justifiable. Behavior rpS1, containing a clear pointing, despite an uncoordinated eye gaze, was misclassified as rpS0 83.33% of the time, which may suggest that this particular behavior would have to be redesigned and made clearer. For rpS2, the results were much more spread, and we hypothesize that the source of misclassifications is a combination of lack of rigorousness on the participants' part as well as low legibility of gaze on the robot's part.

The misclassifications for rN0 came from the same expert, which may suggest that in this case the source of confusion was not from the robot, but from her low experience level with ADOS-2. The low accuracy of rN2 is most probably due to the difficulty in assessing the gaze direction of the robot, as it seemed to be easily confused with rN1, whose main difference is the direction and duration of gaze. For the language feature (rlS), it seems like rlS3 (echolalia, which is easily identifiable) was the only behavior that was immune to misclassifications, while the other three behaviors seem to have been somehow affected by the factors discussed above.

Table 2: Accuracy results per participant across the two studies.

Expert	Total Accuracy (%)			ADOS-2 training	Real-world experience
	Video	Real	Both		
1	95	63	78	Non-official	Low
2	88	75	81	Official	High
3	69	69	69	Official	Very Low

Table 3: Accuracy and agreement results per feature, including comparison with percent agreement values from the ADOS-2 literature with children in ideal (Lord et al. [5]) and naturalistic settings (Zanger et al. [38]).

Features	Accuracy (%)			Agreement (%)					
				Spearman's rho			Percent agreement		
	Video	Real	Both	Video	Real	Both	Video	Children (naturalistic) [38]	Children (ideal) [5]
rN	75	75	75	91	85	80	50	76	84
rJA	100	83	92	100	76	86	100	78	96
rlS	92	75	82	97	91	92	83	80	96
rpS	67	42	54	93	59	76	58	60	85
Combined	**83**	**69**	**76**	**92**	**76**	**83**	**73**	**74**	**90**

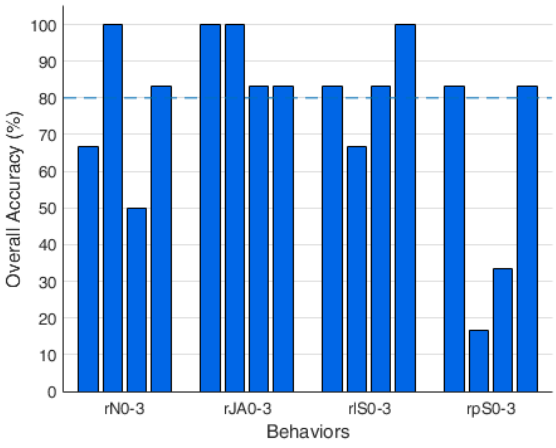

Figure 6: Average accuracies for each robot behavior across the two studies. 11 out of 16 behaviors have an accuracy above 80%.

3.2 Agreement results

As our inter-rater agreement measure, we used both the average Spearman's correlation coefficient to account for the ordinal nature of the data, and the percent agreement for comparison with reference values from studies run with children [5, 38].

Starting with the average Spearman's correlation, we obtained a relatively high agreement value for both studies combined ($\rho = 0.83$). We computed p-values for each pair of raters against the alternative hypothesis that the correlation is greater than zero, using the exact permutation distributions, yielding $p \leq 1.09\text{e}{-7}$ for all pairs of raters, hence indicating general strong agreement between the experts, as expected. Similar to the accuracy results, agreement results differed considerably across the two studies ($\rho = 0.92$ for 'video' vs. $\rho = 0.76$ for 'real'). For both accuracy and agreement results, it is unclear if these differences were mainly due to the embodiment factor, or if the different grouping of behaviors (blocks with the same task versus blocks with the same robot customization) played a role. Looking at individual features, the feature with the highest agreement was the language feature (rlS) ($\rho = 0.92$), which is expected given its high-objectivity coding scheme. The lowest agreement was for the pointing feature (rpS) ($\rho = 0.76$), which also showed a surprisingly large difference in agreement between the video and real

scenarios. We attribute this difference to the same reasons that may have affected accuracy.

The percentage agreement yielded lower or equal values as compared to the previous metric, as expected, since it considers all mismatches have the same weight, achieving an overall value of 72.75%. In the last two columns of Table 3, we report, for comparison, the same metric values from two different sources of the ADOS-2 literature with children. The last column reports values from the ADOS-2 Module 2 manual by Lord et al. [5], obtained from 'research-reliable' ADOS-2 therapists under ideal conditions. The penultimate column reports values obtained in a more naturalistic setting by Zander et al., from clinically trained ADOS-2 users pertaining to 13 different clinical sites [38]. We achieve an agreement similar to the naturalistic setting case as reported by Zander et al. ($PA = 73\%$ vs. $PA = 74\%$), while the ideal setting case [5] shows much larger values ($PA = 90\%$). This result suggests that the sources of disagreement in our solution may be largely due to the common problem of rater subjectivity for non research-reliable ADOS-2 users, hence supporting the objectivity of the ADOS-2 instrument for evaluating robot behavior.

3.3 Order effects

An additional hypothesis on misclassifications is that the participants may have gotten fatigued as the studies progressed. If this were the case, we would expect a positive correlation between the presence of errors and the index at which the behaviors appeared in the study, given the fact that counterbalancing was used. To test this hypothesis, we computed the Spearman's correlation coefficient between those two variables with a single-tailed t-test for statistical significance. We found a statistically significant positive correlation ($\rho = 0.33$, $p = 0.011$) in the video-based study, which suggests that participants were getting fatigued as the survey progressed, making them prone to less sharp judgment. However, interestingly, this effect was not observed with the real robot ($\rho = 0.06$, $p = 0.336$), which we may attribute to the fact that the interaction was more engaging than answering an online survey. Also, the physical presence of the examiner and the learning effects may have contributed to that difference.

3.4 Questionnaire results

We now report the results on the Likert-scale responses from our questionnaire, summarized in Figure 7. Since each item only had three responses, statistical tests will not be used in our analysis, however comparing the mean responses on different items may be indicative of expert opinion and is useful for directing future research endeavors in this space.

Overall, the participants provided high ratings for all three applications we suggested. It is interesting to see that, even though they had previously seen videos of our robot in the first study, their ratings on suitability for therapist training as well as therapy, increased after they had actually interacted with our robot.

In the particular application of therapist training, on average, our solution was rated higher than existing therapist training methods along the dimensions investigated. As expected, our solution was rated as much more interactive than existing solutions. It was also rated as similar in terms of profile diversity and lower in terms of behavior diversity, which is expected since our current prototype only considered three tasks and a single behavior for each feature severity, when in reality many different behaviors may fall under the same category.

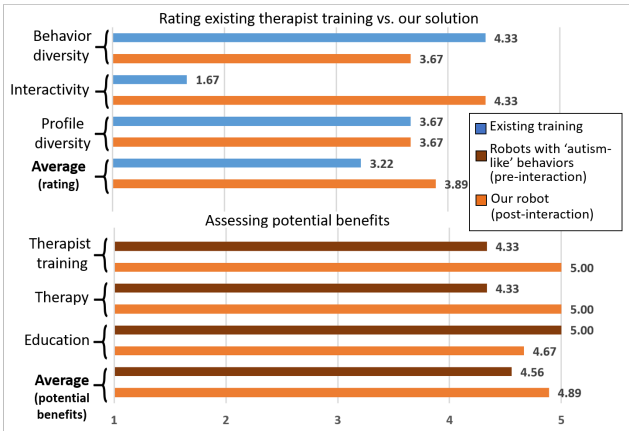

Figure 7: Summary of mean responses to questionnaire items.

3.5 Qualitative observations

Even though at first we felt some skepticism from our participants towards our idea of 'simulating' children with ASD, after both studies, they seemed to be pleasantly surprised by how useful this robotic tool could be, emphasizing features of the robot that they hadn't foreseen.

They understood that what we were trying to replicate was not complex underlying cognitive mechanisms of children with ASD, still poorly understood by scientists, but rather high-level behaviors that are clearly laid out and categorized in available diagnostic tools.

After the video-based survey, expert 2 said she was surprised that the interactions shown in the videos felt "just like ADOS-2 tasks with real kids." After the 'in situ' study, she stressed that what she found very interesting and useful was that she could repeat the same stimulus and observe the same response as many times as she wanted, which she thought made the system particularly useful for therapists in training.

Expert 1 mentioned that she was having trouble keeping up with the interaction, especially when it came to assessing the gaze direction of the robot. This difficulty may explain that she had the lowest accuracy in the 'in situ' study, but is unclear if it was because of the robot or because her level of experience was low (she had actually never performed a real entire session with a child). This observation motivates the importance of the interactive component needed for training, which our solution attempts to address.

3.6 Additional remarks

Perhaps the strongest limitation of our methodology was the small sample size, as well as the lack of ADOS-2 research-certified (as opposed clinically-certified) participants, which we expect would be have increased the reliability of our results. Unfortunately ADOS-2 research-certified professionals are scarce, and finding such individuals to physically interact with our robot was a major challenge. Additionally and from a technological point of view, our robot, being autonomous, showed some variability in some behaviors, which may have injected additional noise in our data. Depending on the intended use, the level of autonomy may need to be adjusted depending on the advantages it provides. Finally, in a real world setting, it may be desirable to sample the combination of feature severities to match the statistical distribution of real ADOS-2 data. In our previous work [39], we devised a data-driven method for sampling such features, which could be interfaced with the robotic solution presented in this paper.

4 Conclusions

We demonstrated our approach on enabling a humanoid robot to exhibit model-based 'autism-like' behaviors of varying severities. We designed 16 behaviors for a NAO robot, following the standardized categorization of the ADOS-2, the gold standard of autism diagnosis. Our behaviors spanned different levels of severity along 4 selected features from the ADOS-2 model. We integrated those behaviors into an autonomous agent running on the robot, hence enabling flexible and continuous interactions with humans. Finally, we evaluated our designed behaviors by running a video-based and an 'in situ' study with three trained ASD therapists.

Our results generally show satisfactory levels of accuracy and agreement for most behaviors, although some behaviors may have to be redesigned to reduce the level of subjectivity in coding some robot motions and poses. In particular, estimating gaze direction appeared to be a challenging component of the robot's behaviors. Despite the systematic coding structure of ADOS-2, we observed considerable levels of subjectivity in coding for some features. This subjectivity is a known problem in behavior-based diagnostic procedures in general [40]. Moreover, as compared to the video-based study, both accuracy and agreement dropped in the real interaction, even though the behaviors of the robot were largely identical. This seems to suggest that the cognitive load of embodied interaction affects the performance of the therapists. These observations therefore motivate the potential use of our solution for complementing therapist training, which currently heavily relies on watching videos. Because current robots can only mimic human behavior in a shallow, exaggerated and simplistic way, an interactive robot capable of simulating simplified versions of a real ADOS-2 interaction may specifically focus on procedural training, as opposed to coding training, for which videos are more adequate.

Our questionnaire results suggest that autism experts are willing to use robotic tools in their professional fields, and holds promise for the use of robots to assist them in their training and practice. The applications we foresee and which were looked at in this research were: complementing therapist training, unlocking novel autism tasks involving robots, and providing interactive tools to educate and sensitize the general population about the diversity of the behavioral aspects of ASD.

Interactive robots exhibiting 'autism-like' behaviors with different severities open the door to a number of exciting applications to train, treat or educate a wide range of individuals dealing with ASD. In future work, we plan

to expand the existing prototype, as well as gather more evidence for the intended uses of our robotic tool. We also plan to apply the same methodology in domains outside of autism, where systematic procedures and models of human behaviors are being utilized to characterize patients, users, clients, or students. We believe that robots that emulate a scale of human behavioral characteristics may unlock many possibilities to create simulated environments as a preparation for critical real-world tasks.

Acknowledgement: This research is partially supported by the CMUP-ERI/HCI/0051/2013 grant, associated with the CMU-Portugal INSIDE project (http://www.project-inside.pt/), as well as national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019. The views and conclusions contained in this document are those of the authors only.

We thank Jocelyn Huang and Patrick Lin for their collaboration in the development of the robot behaviors, as well as Marta Couto and Patrícia Alves-Oliveira for their help. Special thanks go to the Child Development Center at Hospital Garcia de Horta, Almada, Portugal and Liz Carter, Aaron Steinfeld, and Laura Herlant for their feedback.

References

[1] American Psychiatric Association, Diagnostic and statistical manual of mental disorders (DSM-5®), American Psychiatric Publishing, 2013

[2] M. Sigman, S. J. Spence, A. T. Wang, Autism from developmental and neuropsychological perspectives, Annual Review of Clinical Psychology, 2006, 2, 327-355

[3] R. Muhle, S. V. Trentacoste, I. Rapin, The genetics of autism, Pediatrics, 2004, 113(5), e472-e486

[4] P. Chaste, M. Leboyer, Autism risk factors: genes, environment, and gene-environment interactions, Dialogues in Clinical Neuroscience, 2012, 14(3), 281-292

[5] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, S. Bishop, Autism diagnostic observation schedule – Second edition (ADOS-2), Los Angeles: Western Psychological Services, 2012

[6] K. Baraka, F. S. Melo, M. Veloso, 'Autistic robots' for embodied emulation of behaviors typically seen in children with different autism severities, In: Proceedings of the 2017 International Conference on Social Robotics, Springer, Cham, 2017, 105-114

[7] H. Van Ditmarsch, W. Labuschagne, My beliefs about your beliefs: a case study in theory of mind and epistemic logic, Synthese, 2007, 155(2) 191-209

[8] F. Dinum, R. Prada, G. J. Hofstede, From autistic to social agents, In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2014, 1161-1164

[9] K. Richardson, M. Coeckelbergh, K. Wakunuma, E. Billing, T. Ziemke, P. Gomez, B. Vanderborght, T. Belpaeme, Robot enhanced therapy for children with autism (DREAM): A social model of autism, IEEE Technology and Society Magazine, 2018, 37(1), 30-39

[10] B. Scassellati, H. Admoni, M. Matarić, Robots for use in autism research, Annual Review of Biomedical Engineering, 2012, 14, 275-294

[11] K. Dautenhahn, I. Werry, Towards interactive robots in autism therapy: background, motivation and challenges, Pragmatics and Cognition, 2004, 12(1), 1-35

[12] J. Psotka, Immersive training systems: Virtual reality and education and training, Instructional Science, 1995, 23(5-6), 405-431

[13] P. Dieckmann, D. Gaba, M. Rall, Deepening the theoretical foundations of patient simulation as social practice, Simulation in Healthcare, 2007, 2(3), 183-193

[14] A. T. Lee, Flight simulation: virtual environments in aviation, Routledge, 2017

[15] N. E. Seymour, VR to OR: a review of the evidence that virtual reality simulation improves operating room performance, World Journal of Surgery, 2008, 32(2), 182-188

[16] M. J. Shapiro, J. C. Morey, S. D. Small, V. Langford, C. J. Kaylor, L. Jagminas, S. Suner, M. L. Salisbury, R. Simon, G. D. Jay, Simulation based teamwork training for emergency department staff: does it improve clinical team performance when added to an existing didactic teamwork curriculum?, BMJ Quality and Safety, 2004, 13(6), 417-421

[17] M. Macedonia, Games, simulation, and the military education dilemma, In: Internet and the University: 2001 Forum, Louisville, CO, Educause, 2002, 157-167

[18] R. Querrec, C. Buche, E. Maffre, P. Chevaillier, SécuRéVi: virtual environments for fire-fighting training, In: Proceedings of the 5th Virtual Reality International Conference, 2003, 169-175

[19] S. E. Kirkley, J. R. Kirkley, Creating next generation blended learning environments using mixed reality, video games and simulations, TechTrends, 2005, 49(3), 42-53

[20] M. Baptista, C. R. Martinho, F. Lima, P. A. Santos, H. Prendinger, Improving learning in business simulations with an agent-based approach, Journal of Artificial Societies and Social Simulation, 2014, 17(3), 7

[21] H. Lin, C. T. Sun, Problems in simulating social reality: Observations on a MUD construction, Simulation and Gaming, 2003, 34(1), 69-88

[22] S. Babu, E. Suma, T. Barnes, L. F. Hodges, Can immersive virtual humans teach social conversational protocols?, In: Transactions of Virtual Reality Conference, IEEE, 2007, 215-218

[23] M. J. Smith, E. J. Ginger, K. Wright, M. A. Wright, J. L. Taylor, L. B. Humm, D. E. Olsen, M. D. Bell, M. F. Fleming, Virtual reality job interview training in adults with autism spectrum disorder, Journal of Autism and Developmental Disorders, 2014, 44(10), 2450-2463

[24] J. Rickel, W. L. Johnson, Animated agents for procedural training in virtual reality: Perception, cognition, and motor control, Applied Artificial Intelligence, 1999, 13(4-5), 343-382

[25] K. Fast, T. Gifford, R. Yancey, Virtual training for welding, In: the Third IEEE and ACM International Symposium on Proceedings of Mixed and Augmented Reality, IEEE, 2004, 298-299

[26] N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: a three-factor theory of anthropomorphism, Psychological review, American Psychological Association, 2007, 114(4), 864

[27] B. Ingersoll, The social role of imitation in autism: Implications for the treatment of imitation deficits, Infants and Young Children, 2008, 21(2), 107-119

[28] G. Dawson, A. Adams, Imitation and social responsiveness in autistic children, Journal of Abnormal Child Psychology, 1984, 12(2), 209-226

[29] S. Lemaignan, A. Jacq, D. Hood, F. Garcia, A. Paiva, P. Dillenbourg, Learning by teaching a robot: The case of handwriting, IEEE Robotics and Automation Magazine, 2016, 23(2), 56-66

[30] P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, D. Piepol, Building interactive virtual humans for training environments, In: Proceedings of I/ITSEC, 2007, 174, 911-916

[31] S. Rossi, F. Ferland, A. Tapus, User profiling and behavioral adaptation for HRI: a survey, Pattern Recognition Letters, 2017, 99, 3-12

[32] A. Tapus, C. Ţăpuş, M. J. Matarić, User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy, Intelligent Service Robotics, 2008, 1(2), 169

[33] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, N. Hagita, Adapting robot behavior for human–robot interaction, IEEE Transactions on Robotics, 2008, 24(4), 911-916

[34] J. C. Gillesen, E. I. Barakova, B. E. Huskens, L. M. Feijs, From training to robot behavior: Towards custom scenarios for robotics in training programs for ASD, In: Transactions of the 2011 IEEE International Conference on Rehabilitation Robotics, IEEE, 2011, 1-7

[35] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A. S. Rusu, D. O. David, Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments, Interaction Studies, 2012, 13(3), 315-347

[36] E. I. Barakova, W. Chonnaparamutt, Timing sensory integration – robot simulation of autistic behavior, IEEE Robotics & Automation Magazine, 2009, 16(3), 51

[37] B. Scassellati, Investigating models of social development using a humanoid robot, In: B. Webb, T. Consi (Eds.), Biorobotics, MIT Press, 2000

[38] E. Zander, C. Willfors, S. Berggren, N. Choque-Olsson, C. Coco, A. Elmund et al., The objectivity of the Autism Diagnostic Observation Schedule (ADOS) in naturalistic clinical settings, European Child and Adolescent Psychiatry, 2016, 25(7), 769-780

[39] K. Baraka, F. S. Melo, M. Veloso, Data-driven generation of synthetic behavioral feature vectors modeling children with autism spectrum disorders, In: Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, 2017, IEEE

[40] T. Falkmer, K. Anderson, M. Falkmer, C. Horlin, Diagnostic procedures in autism spectrum disorders: a systematic literature review, European Child and Adolescent Psychiatry, 2013, 22(6), 329-340