Design, Development, and Evaluation of a Noninvasive Autonomous Robot-Mediated Joint Attention Intervention System for Young Children With ASD

Zhi Zheng, *Member, IEEE*, Huan Zhao, Amy R. Swanson, Amy S. Weitlauf, Zachary E. Warren, and Nilanjan Sarkar, *Senior Member, IEEE*

***Abstract*—Research indicates that human–robot interaction can help children with autism spectrum disorder (ASD). While most early robot-mediated interaction studies were based on free interactions, recent studies have shown that robot-mediated interventions that focus on the core impairments of ASD such as joint attention deficit tend to produce better outcomes. Joint attention impairment is one of the core deficits in ASD that has an important impact in the neuropsychological development of these children. In this paper, we propose a novel joint attention intervention system for children with ASD that overcomes several existing limitations in this domain such as the need to use body-worn sensors, nonautonomous robot operation requiring human involvement and lack of a formal model for robot-mediated joint attention interaction. We present a fully autonomous robotic system, called noncontact-responsive robot-mediated intervention system, that can infer attention through a distributed noncontact gaze inference mechanism with an embedded least-to-most (LTM) robot-mediated interaction model to address the current limitations. The system was tested in a multisession user study with 14 young children with ASD. The results showed that participants' joint attention skills improved significantly, their interest in the robot remained consistent throughout the sessions, and the LTM interaction model was effective in promoting the children's performance.**

***Index Terms*—Children with autism spectrum disorder (ASD), joint attention, robot-assisted intervention.**

I. Introduction

HUMAN–ROBOT interaction (HRI), in recent years, has been investigated as a potential intervention tool for children with autism spectrum disorder (ASD) [1]. ASD is a common neurodevelopmental disorder that impacts 1 in 68 children in the U.S. [2]. Social communication deficits are among the core impairments of ASD [3]. Robot-mediated interventions are promising in teaching social communication skills to children with ASD because many children with ASD are fascinated by technology and may pay more attention to a robot rather than a human therapist [4]. A few experimental studies have shown that robots could elicit social communication behaviors, such as speech [5] and arm gesture imitation [6], better than humans within the studies. In addition, robots are highly controllable, precise, and could potentially provide effective interventions with low cost [7], [8].

Primarily animal-like robots [9], [10] and small humanoid robots [11], [12] have been used for studies with children with ASD. Kozima *et al.* [13] designed a small creature-like robot called "Keepon," which successfully elicited positive social interaction behaviors in children with ASD. A humanoid robot called "KASPAR" [14], [15] was successfully used to facilitate collaborative play and tactile interaction with children with ASD. Feil-Seifer and Matarić [16] found that contingent activation of a robot during interactions yielded immediate short-term improvement in social interactions. While important to demonstrate the potential of HRI in ASD intervention, most of these earlier studies chose free play as the mode of interaction instead of focusing on the core deficits of ASD. However, studies in ASD intervention have shown that interventions are most effective when the intervention is focused on the core deficits of ASD [17]. In addition, most of these previous HRI systems were open-loop systems and thus were not responsive to the dynamic interaction cues from the participants to be able to adapt and individualize intervention. The primary goal of the current work is to design a fully autonomous closed-loop robotic system that can target core deficits of ASD. Targeting core deficits

Manuscript received September 14, 2016; revised April 11, 2017 and August 17, 2017; accepted October 25, 2017. Date of publication December 14, 2017; date of current version March 13, 2018. This work was supported in part by Vanderbilt Kennedy Center (Hobbs Grant), in part by the National Science Foundation under Grant 1264462, in part by the National Institutes of Health under Grant 1R01MH091102-01A1 and Grant 1R21MH103518-01, in part by the National Institute of Child Health and Human Development under Grant P30HD15052, and in part by the National Center for Advancing Translational Sciences under Grant UL1TR000445-06. This paper was recommended by Associate Editor Dr. Zhelong Wang. *(Corresponding author: Zhi Zheng.)*

Z. Zheng is with the Department of Biomedical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI 53212 USA (e-mail: zheng36@uwm.edu).

H. Zhao is with the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37212 USA (e-mail: huan.zhao@vanderbilt.edu).

A. R. Swanson, A. S. Weitlauf, and Z. E. Warren are with the Treatment and Research in Autism Disorder, Vanderbilt University Kennedy Center, Nashville, TN 37203 USA (e-mail: amy.r.swanson@vanderbilt.edu; amy.s.weitlauf@vanderbilt.edu; zachary.e.warren@vanderbilt.edu).

N. Sarkar is with the Department of Mechanical Engineering, Vanderbilt University, Nashville, TN 37212 USA (e-mail: nilanjan.sarkar@vanderbilt.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2017.2776865

using robot-mediated intervention is more complex since the autonomous system needs to elicit response regarding the core deficit through a set of well-designed interaction protocol, assess participant's response in real time, and adapt its (i.e., the robot's) own interaction to shape the participant's response.

We introduce a new closed-loop fully autonomous robotic system, named NORRIS, which stands for *No*ncontact-*R*esponsive *R*obot-mediated *I*ntervention *S*ystem, to help children with ASD learn joint attention skills. Joint attention is the process of sharing attention and socially coordinating attention with others to effectively learn from the environment [18]. Joint attention skills underlie the neurodevelopmental cascade of ASD, and successful intervention targeted on joint attention is essential to improve numerous other developmental skills in children with ASD [19]–[21].

The current work improved our previous work [22], [23] in a number of important ways. While Bekele *et al.* [22] presented a novel HRI architecture for ASD intervention, adaptive robot-mediated intervention architecture (ARIA), and developed an effective least-to-most (LTM) protocol for joint attention intervention with promising results, it required participants to wear an instrumented hat for gaze inference. Since many young children with ASD are sensitive to unfamiliar touch [24], close to 40% children did not want to wear the hat and thus could not take part in the intervention. In order to solve this problem, Zheng *et al*. [23] developed another robotic system that inherited the LTM protocol from ARIA but used a Wizard of Oz [25] strategy for gaze detection to eliminate the need for the instrumented hat. While it enabled 100% participation, the system became semiautonomous and needed human involvement for gaze inference. Additionally, both studies in [22] and [23] used LTM protocol for joint attention but did not provide a generalizable mathematical model for LTM interaction. In LTM, the teacher allows the learner an opportunity to respond independently on each training stage and delivers the least intrusive prompt first. If necessary, more intrusive prompts, usually upgraded based on the previous prompts, are then delivered to the learner to complete each training procedure [26]. Essentially, LTM provides support to the learner only when needed. LTM has been widely applied in diagnostic and screening tools for children with ASD [27], [28]. However, to our knowledge, no mathematical model of LTM has been presented in the literature such that LTM-based interaction can be generalized for multiple skill training. Anzalone *et al*. [29] have developed a joint attention intervention system using a robot administrator and external cameras to sense the attention of the participants, which is similar to the setup presented in NORRIS. However, NORRIS introduces novel system architecture, gaze-tracking method, LTM interaction mathematical model, and multisession user study for young children with ASD that have not been reported in the literature.

The contributions of the current study are twofold: 1) development of a new fully autonomous closed-loop robot-mediated intervention system that can infer gaze noninvasively and is capable of administering LTM protocol based on a general mathematical model; and 2) results from a feasibility joint attention intervention user study that tested the newly developed system in a multisession study. This study validated the effectiveness of LTM in the robot-mediated intervention, i.e., adding prompt levels increased the probability of expected response and the system elicited expected response with high probability at the highest prompt level.

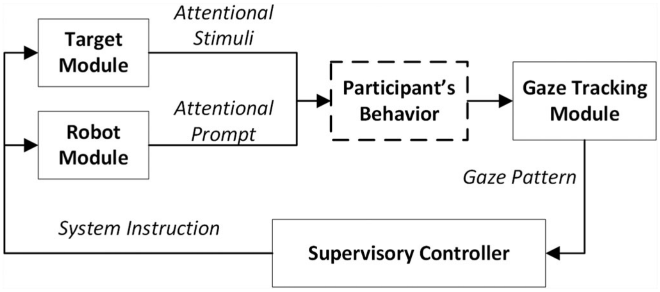


Fig. 1. NORRIS system architecture.

The remainder of this paper is organized as follows: Section II describes the architecture and components of NORRIS. Section III introduces the mathematical model for LTM as the interaction logic. Section IV describes the design of the feasibility multi-session user study to validate the NORRIS system as well as the results of this user study. Section V presents the summary of contributions and limitations of the paper, respectively.

II. NORRIS System Architecture and Components

A. System Architecture

HRI using NORRIS is designed to work as follows. A child with ASD will be seated in a room in front of a humanoid robot. The room will be equipped with a set of spatially distributed computer monitors or TVs where audio–visual stimuli will be presented. The robot will administer an LTM-based joint attention prompting protocol to the child and the child's response in terms of gaze direction will be inferred by a set of distributed cameras. Based on whether the child shares attention or not, the robot will provide appropriate feedback and move on to the next prompt. As shown in Fig. 1, NORRIS has four main components.

1) The robot module controls robot actions.
2) The target module controls environmental factors.
3) The gaze-tracking module provides interaction cue sensing.
4) The supervisory controller controls the interaction logic.

The supervisory controller is the "brain" of NORRIS that sends commands to the robot and the target module to present directional prompts to the participant. For example, the robot can turn its head to a monitor displaying a picture, and ask the participant to look at that monitor. The participant may or may not look at the monitor, and this looking behavior is sensed by the gaze-tracking module. The tracking module further computes whether the direction of the participant's gaze falls on the monitor, and sends this message back to the supervisory controller. Then, the supervisory controller sends commands to the robot and the target module again telling what to show next,

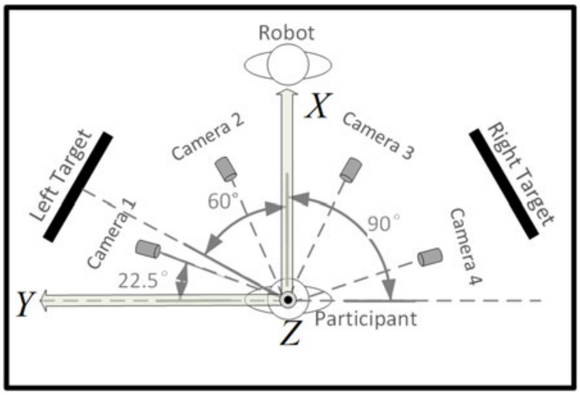


Fig. 2. Top view of NORRIS and the global frame.

based on an interaction protocol. Therefore, NORRIS provides a fully autonomous closed-loop interaction between the system and the participant.

B. System Components

1) Robot Module: A humanoid robot NAO by Aldebaran Robotics [30] was embedded in NORRIS. NAO has been widely applied for children with ASD [11], [22] due to its attractive childlike appearance and high controllability. We designed a new controller for NAO that communicates with the supervisory controller. The robot controller was embedded with a built-in library storing all the necessary motions (e.g., turning its head to a monitor) and speeches (e.g., asking the participant to look toward a target) needed for the interaction. The robot's actions are detailed in Section III-B along with the interaction protocol.

2) Target Module: Two flat TVs (width: 70 cm; height: 43 cm), one to the right and one to the left of the participant, were used as attentional targets. The robot would point to one of the monitors at a time and ask the participant to look at what was being shown on that monitor. The two TVs were controlled individually by two target controllers that received commands from the supervisory controller. A library of pictures, audios, and videos were embedded in the target controller. Based on the commands sent from the supervisory controller, static pictures, audios, or videos of children's interest were displayed. The set of target actions are detailed in the interaction protocol (see Section III-B).

3) Gaze-Tracking Module: The gaze-tracking module detected the participants' looking behavior. The direction of a participant's gaze was computed based on the orientation of his/her head as detected by a set of cameras as shown in Figs. 2 and 3. Fig. 2 illustrates the top view of NORRIS in the global reference frame. The center of the participant's head was the origin of the global frame. The *X*-axis and the *Y*-axis pointed forward and to the left of the participant, respectively, and the *Z*-axis pointed upward out of the plane. Fig. 3 shows both the body-attached head frame of the participant and the global reference frame that share the same origin. If the participant did not perform yaw (around the *Z*-axis), pitch (around the *Y*-axis), or roll rotations (around the *X*-axis), the head frame was aligned with the global frame. The unit vector along the positive *x*-axis of the head frame, $\overrightarrow{V_f}$, represents the frontal head orientation, and was used to derive the gaze direction. Four cameras were employed for gaze detection, each with its own coordinate system. The gaze-tracking method has three steps as discussed next. It is to be noted that while Steps 1 and 2 were inherited from our previous work [31], Step 3 was newly developed in the current study.

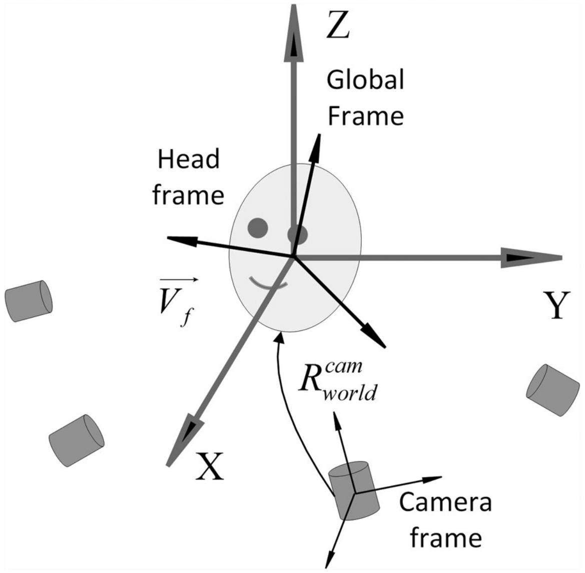
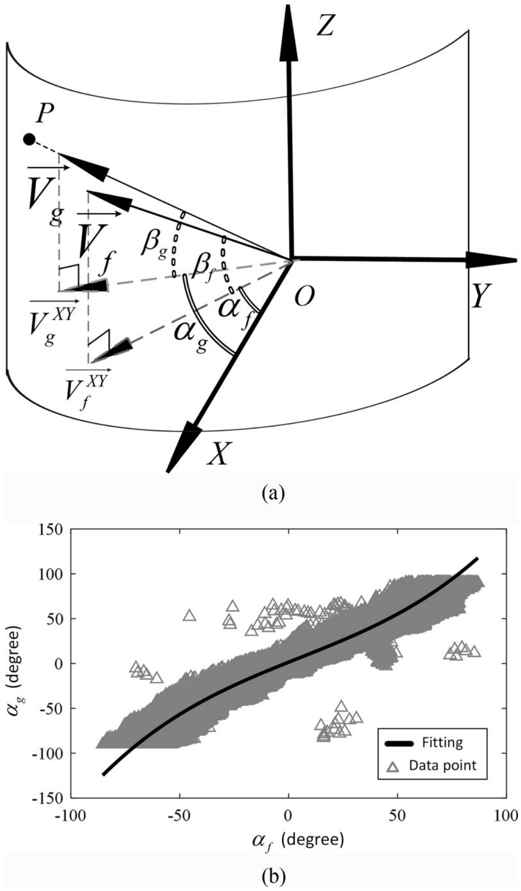


Fig. 3. Coordinate systems of gaze tracking.

Step 1: Detect head orientation from a camera. The supervised descent method [32] method was applied to each camera to achieve fast and robust head orientation estimation with respect to the camera's frame. The image of the participant's frontal face is needed for this estimation, and thus a given camera can only detect head orientation when the frontal face is visible to it. The detected head orientation is represented by $\overrightarrow{V_f}$ in the camera's frame. However, in the current joint attention study, we wanted to detect a larger head yaw angle (about 180°) than what can be detected by one camera (about 80°) for realistic tasks. Therefore, we developed a distributed head orientation estimation algorithm for an array of four cameras (as shown in Fig. 2) around the participant with partially overlapping views to extend the detection range. This design guaranteed that no matter which part of the interaction environment the participant was looking at, at least one camera could capture his/her frontal face in order to conduct head orientation estimation.

Step 2: Transform the head orientation estimation from a camera's frame to the global frame. Each camera was calibrated to get the transformation matrix, R^{cam}_{world}, between the camera's frame and the global frame. As shown in Fig. 3, $R^{cam}_{world} R^{cam}_{world}$ transform $\overrightarrow{V_f}$ from the camera's frame to the global frame.

Step 3: Compute gaze direction $\overrightarrow{V_g}$ from $\overrightarrow{V_f}$ in the global frame. As shown in Fig. 4(a), $\overrightarrow{V_g}$ denotes the gaze direction in the global frame. We used the vertical and horizontal components of $\overrightarrow{V_g}$ to judge whether the participant's gaze direction fell



(a)

(b)

Fig. 4. (a) Gaze direction computation in the global frame. (b) Illustration of $\overrightarrow{V_f}$, $\overrightarrow{V_f^{XY}}$, α_f, β_f, $\overrightarrow{V_g}$, $\overrightarrow{V_g^{XY}}$, α_g, andβ_g. Mapping from horizontal head orientation to horizontal gaze direction.

into a center range (e.g., the range of a target monitor). In Fig. 4, $\overrightarrow{V_g^{XY}}$ is the projection of $\overrightarrow{V_g}$ on the *XY* plane. The angle between $\overrightarrow{V_g}$ and $\overrightarrow{V_g^{XY}}$, β_g, denotes the vertical gaze direction. The angle between the *X*-axis and $\overrightarrow{V_g^{XY}}$, α_g, denotes the horizontal gaze direction. $\overrightarrow{V_f^{XY}}$ is the projection of $\overrightarrow{V_f}$ on the *XY* plane. The angle between $\overrightarrow{V_f}$ and $\overrightarrow{V_f^{XY}}$, β_f, is used for computing β_g. The angle between the *X*-axis and $\overrightarrow{V_f^{XY}}$, α_f, is used for computing α_g.

In order to correlate head orientation with gaze direction, we conducted a small study with ten adults where the volunteers were asked to look at a marker in front of them that was moved from left to right five times followed by a right to left marker movement for another five times.

In the horizontal direction, we trained a mapping function to derive a_g from a_f. a_g was identified as the angle of the moving marker [between −90° (left) and 90° (right)]. Simultaneously, the volunteers' horizontal head orientation, a_f, was estimated by the camera array. A total of 22 236 data pairs were collected. We used a polynomial fitting to reflect the relation between a_g and a_f. In Fig. 4(b), the blue points indicate the pairs (a_f, a_g). The red curve with a sigmoid shape is the curve that maps α_f. The equation of the red curve is

$$\alpha_g = 5.88^{-5}\alpha_f{}^3 - 8.94^{-4}\alpha_f{}^2 + 0.97\alpha_f + 1.50. \quad (1)$$

The mean distance from the data points to the red curve is 9.12°. We can see that, in general, a larger α_f leads to a larger $\|\alpha_g - \alpha_f\|$. Intuitively, the more the participant's head turned to the side, the larger the deviation between the gaze direction and the frontal head orientation in the horizontal direction.

The vertical gaze direction was approximated by $\beta_g = \beta_f - \beta_{\text{baseline}}$. Here, $\beta_{baseline}$ is an offset angle which was calibrated for each participant. During the calibration, the participant was prompted to look along the *X*-axis, and β_f at this moment was recorded as $\beta_{baseline}$. In the user study presented in Section IV, $\beta_{baseline}$ β_{baseline} ranges from –11.57° to 8.14°.

The range of both monitors can be represented by the values α_g and β_g. The α_g range of the left and the right monitors was [–70°, –50°] and [50°, 70°], respectively. However, in order to accommodate for mapping error as well as encouraging young children with ASD to continue with the intervention, we relaxed the range by 15° on each side. Similarly, the range of β_g was [−26.3°, 12.5°] from top to bottom, which covered an additional 43 cm (the height of the monitor) beyond the monitor's top and bottom edges. Therefore, if $\alpha_g \in [-85°, -35°]$, and $\beta_g \in [-26.3°, 12.5°]$, the system would infer that the participant responded to the left monitor. Similar ranges were applied for the right monitor. On average, the whole gaze-tracking module refreshed at a speed of 15 fps. Note that even though this gaze-tracking method was developed for the NORRIS system in this study, it is an independent component which can be easily applied to other scenarios where large range gaze tracking is needed.

4) Supervisory Controller: The supervisory controller communicated with different system components and controlled the global logic of the interaction. The communication was implemented with the transmission control protocol/Internet protocol socket communication method. The average communication time from sending a message to receiving the message between the supervisory controller and a system component was about 25 ms, which guaranteed real-time closed-loop interaction. The global interaction logic, which we call the interaction protocol, is discussed in detail in Section III.

III. LTM Interaction Protocol

The LTM hierarchy was applied in NORRIS to form the interaction protocol. LTM has been widely applied in diagnostic and screening tools for children with ASD [27], [28]. In LTM, the teacher allows the learner an opportunity to respond independently on each training stage and delivers the least intrusive prompt first. If necessary, more intrusive prompts, usually upgraded based on the previous prompts, are then delivered to the learner to complete each training procedure [26]. Essentially, LTM provides support to the learner only when needed.

LTM has been applied to a few important robot-mediated intervention systems for children with ASD. Feil-Seifer and Matarić [33] and Greczek *et al.* [11] introduced a graded cueing feedback mechanism to teach imitation skills to children with

ASD. In this mechanism, higher prompts were upgraded based on the initial prompt by adding additional verbal and gestural hints to help children copy gestures. Huskens *et al*. [34] used a robot to prompt question-asking behaviors in children with ASD. The robot used open-question prompt initially. If the participants did not respond correctly, the robot would add more hints (e.g., adding part of the correct response) in the following prompts. Zheng *et al*. [35] designed a robot-mediated imitation learning system using a prompting protocol to help address an incorrect imitation. The robot first showed a gesture to the participant and asked him/her to copy it. If the child could not do it correctly, the robot would point out where to improve in the following prompts. Bekele *et al*. [22] developed the ARIA system to teach joint attention skills to children with ASD. If the participants did not respond to simple directional prompts given by a robot, higher levels of prompts with additional visual and verbal directional hints were provided. Kim *et al*. [36] designed a robot-assisted pivotal response training platform, where the higher levels of prompts were built by adding target responses hints on lower level of prompts.

From these examples, we can see that LTM is not limited to a specific skill, but can be used as a general guidance mechanism for robot-mediated intervention. However, to our knowledge, no mathematical model has been proposed to create a general LTM framework. In this work, for the first time, we attempt to develop a general LTM-based robot-mediated intervention (LTM-RI) model. Such a model can be used to teach different skills to children with ASD as well as adapt the prompts for a specific skill. We expand the model for joint attention intervention, which is used for the user study.

A. LTM-RI Model

An intervention system uses prompts to teach a skill to children with ASD. These prompts may consist of robot actions (e.g., motions and speeches), and may also include environmental factors that coordinate with the robot's actions (e.g., the attentional target that the robot may refer to). Suppose we have libraries of different robot actions $\{RA\}$ and environmental factors $\{EF\}$, then we can combine different RA_i and EF_j to form different prompts. These combinations may have different strength in eliciting the expected response (e.g., child looking at the target monitor), $\mathrm{ExpResp}$, from a child. For example, in the current joint attention study, the robot turning its head (*RA*) to a static picture display on the TV (*EF*) might have a weaker impact on the children than pointing (*RA*) to a cartoon video displayed on the monitor (*RA*). Here, we arrange the order of the elements in $\{RA_i\}$ and $\{EF_j\}$ as follows: RA_a is stronger (includes more instructive information) than RA_b if a > b; and EF_c is stronger than EF_d if c > d. These orders can be determined based on common sense and clinical experiences.

LTM-RI starts from presenting the weakest combination of *RA* and *EF* to form the least intrusive prompt. If this cannot elicit $\mathrm{ExpResp}$ (expected response), stronger *RAs* and *EFs* will be provided iteratively to form more instructive prompts, until the end of an intervention trial. We formally define this iterative procedure as follows.

```
LTM-RI Model
Step 1: Initial prompt (prompt level = 1).
        $\mathrm{Behavior(1) = BF(RobAction(1), EnviFactor(1))}$
        $\mathrm{Resp(1) = ICD(Behavior(1))}$
        If $\mathrm{Resp(1) = ExpResp}$
            Reward
          Go to Step 3
Step 2: Iterative prompting loop.
          For prompt level n = 2: IN
        $\mathrm{[RobAction(n), EnviFactor(n)] = PF(Resp(n-1))}$
        $\mathrm{Behavior(n) = BF(RobAction(n), EnviFactor(n))}$
        $\mathrm{Resp(n) = ICD(Behavior(n))}$
        If $\mathrm{Resp(n) = ExpResp}$
              Reward
            Break
        $n = n + 1$
Step 3: Termination.
          Robot naturally stops the interaction
```

Here, BF is an implicit function that describes the participant's behavior (e.g., participant's gaze direction) given *RA* (*RobAction)* and *EF (EnviFactor)*. This behavior is sensed by the interactive cue detection function (*ICD*) to determine whether the behavior is *ExpResp*. In the current study, *ICD* is the gaze-tracking module. In the simplest scenario, we can categorize $\mathrm{Resp(n) = ExpResp}$ and $\mathrm{Resp(n) \neq ExpResp}$. *PF* is the prompting function which decides what *RA* and *EF* to present, if$\mathrm{Resp}(n) \neq \mathrm{ExpResp}$. Therefore, *PF* is a sorted list of prompt levels following the LTM heirarchy. We want to identify the lowest level of support needed by the participant to perform$\mathrm{ExpResp}$. If $\mathrm{Resp}(n) \neq Ex\mathrm{pResp}$ (the response detected was not the expected response), given $\mathrm{RobAction(n-1)} = RA_i$ (robot action at time instance *n* – 1), and $\mathrm{EnviFactor(n-1)} = EF_j$ (environmental factor at time instance *n* – 1), we choose the robot action for the next instant, $\mathrm{RobAction(n)} = RA_l (l \geq i)$, from {*RA*} and the environmental factor for the next time instance, $\mathrm{EnviFactor(n)} = \mathrm{EF}_k (k \geq j)$, from {*EF*}, so that the next prompt repeats the last prompt or provides a more instructive prompt. LTM-RI steps works are as follows.

1) In Step 1, the participant's baseline behavior $\mathrm{Perf}(1)$ is evaluated by prompt level 1, which consists of the weakest *RA (*$\mathrm{RobAction}(1)$) and *EF* ($\mathrm{EnviFactor}(1)$). If the participant's response, $\mathrm{Resp}(1)$, is $\mathrm{ExpResp}$, then higher prompts are not needed. The system gives rewards and then executes Step 3 to terminate the intervention. Otherwise, Step 2 is executed.
2) In Step 2, prompt level 2 is given first. If the participant cannot perform $\mathrm{ExpResp}$, the higher prompts are presented one by one until level *IN*. During this iteration, the next level of prompt ($\mathrm{RobAction}(n)$ and $\mathrm{EnviFactor}(n)$) is formed based on the current response of the participant $\mathrm{Resp(n-1)}$, according to the *PF*. If $\mathrm{Resp(n) = ExpResp}$, the system gives a reward to the participant and goes to Step 3.

TABLE I
PROMPT LEVELS

Prompt level	Prompting element list
1 and 2	$RA_1 + EF_1$
3 and 4	$RA_2 + EF_1$
5	$RA_2 + EF_2$
6	$RA_2 + EF_3$
Prompt elements (TR means target monitor)	
RA_1	Robot turned its head to the TM, saying “Look!”
RA_2	Robot turned its head and pointed its arm to the TM, saying “Look over there!”
EF_1	TM displayed a static picture.
EF_2	TM displayed an audio clip.
EF_3	TM displayed a video clip.

We can see that if ExpResp happened on prompt level *n*, it means that level 1 to *n* – 1 have been executed but failed to elicit ExpResp. Suppose ER_n means $\mathrm{Resp(n)} = \mathrm{ExpResp}$, and PT_x represents prompt level *x* had been executed but was not successful. Then, $P(ER_n|PT_1, \ldots, PT_{n-1})$ represents the probability that ExpResp happens on prompt level *n*. In order to measure the impact of LTM-RI, we define an intensity function I_n as follows:

$$I_n = P\left(ER_1\right) + P\left(ER_2|PT_1\right) + \cdots + P\left(ER_n|PT_{n\,-\,1}, \ldots PT_1\right). \tag{2}$$

I_n represents the probability of ExpResp at or before prompt level *n*. LTM-RI procedure has two goals:

Goal 1: $I_m > I_n$, given $m > n$. This means adding prompt levels increases the probability of ExpResp.

Goal 2: $I_{IN} = 1 - \varepsilon$, $\varepsilon = 0$ or ε is a small positive number. This means that eventually, at the highest prompt level, the system can elicit ExpResp with high probability.

We can see that the LTM-RI is a general model that is not limited to one particular skill. What behaviors of the participants that the model tracks depend on the design of *ICD*. The number of prompt levels and the content of the prompts can be easily adjusted within this framework by changing the detail of *PF*. In the current work, LTM-RI is itemized for joint attention intervention below. While the specific content of the prompts in the six-level hierarchy was inherited from our previous studies [22], [23], we interpret it under the framework of LTM-RI to demonstrate how to implement the model.

B. LTM-RI Trial in the Current Study

We designed the intervention trial of NORRIS based on LTM-RI. The *RAs* and *EFs* applied are shown in Table I, which is the *PF* in LTM-RI. Larger the subscript implies stronger directional information.

In Step 1 of LTM-RI (prompt level 1), the robot turned its head to the target monitor, saying “Look!” (RA_1). At the same time, the monitor displayed a static picture (EF_1).

In Step 2 of LTM-RI, IN = 6. Prompt level 2 was the same as prompt level 1. In prompt levels 3 and 4, the robot not only turned its head, but also pointed its arm to the target monitor, saying “Look over there!” (RA_2). At the same time, the monitor still displayed a static picture (EF_1). In prompt levels 5 and 6, the robot action was kept as RA_2, but the monitor displayed an audio clip (EF_2) and a video clip (EF_3), respectively. At any time during a trial, if the participant looked at the target (*ExpResp* happened), the robot would say “Good job!” and the target monitor would display cartoon video for 10 s as rewards. Otherwise, the prompt level would be presented one by one until prompt level 6 was completed.

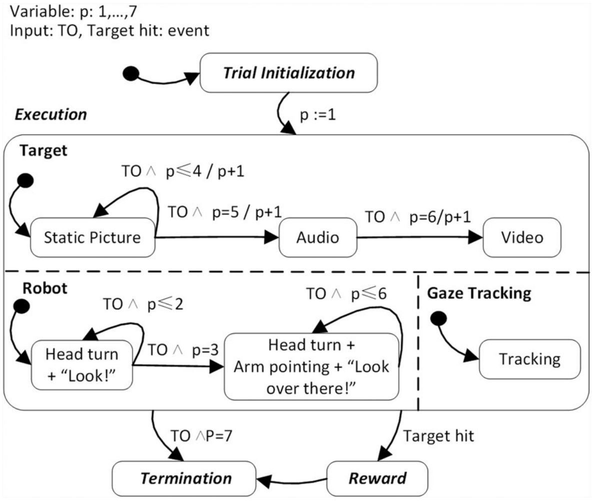


Fig. 5. Harel Statechart model of the NORRIS LTM-RI trial.

Finally, Step 3 of LTM-RI was executed, where the robot returned to its standing position, thanked the participant, and said Goodbye.

This protocol closely aligns with standardized diagnostic assessment procedures (i.e., Autism Diagnostic Observation Schedule–Second Edition [27]) for demarcating joint attention symptoms in young children with ASD as well as robot capacity.

In order to implement LTM-RI within NORRIS, we interpreted the LTM-RI trial with the standard Harel Statechart model [37], which is an extended state machine capable of modeling hierarchical and concurrent system states. As shown in Fig. 5, rectangles denote states. When an event happens, a state transition takes place, which is indicated by a directed arrow. Solid rectangles mark exclusive-or (XOR) states, and the dotted lines mark AND states. Encapsulation represents the hierarchy of the states. In the same hierarchy (encapsulated by the same rectangle), the system must be in only one of its XOR states, while in all of its AND states. Therefore, the AND states represent parallel processes in the system.

The first hierarchy includes four XOR states:

$$\mathrm{S1} = \{\mathrm{Initialization}, \mathrm{Execution}, \mathrm{Reward}, \mathrm{Termination}\}.$$

At the beginning of a trial, the system is in the *Initialization* state, where the robot stands straight facing the participant. Then, the system transits to the *Execution* state and initializes variable *p = 1*.

The *Execution* state is the second hierarchy, which includes three AND states, showing target, robot, and gaze-tracking

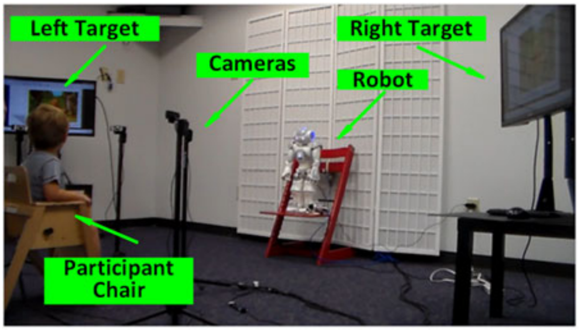


Fig. 6. Experiment room configuration.

modules running in parallel

$$\text{Execution} = \{\text{Target, Robot, Gaze tracking}\}. \quad (3)$$

The third hierarchy controls the prompts

$$\text{Target} = \{\text{Static Picture, Audio, Video}\} \quad (4)$$

$$\begin{aligned}\text{Robot} = \{&\text{Headturn} + \text{“Look!”},\\ &\text{Headturn} + \text{Armpointing} + \text{“Lookoverthere!”}\}.\end{aligned} \quad (5)$$

The *Target* state includes three *EF*s, and the *Robot* state includes two *RA*s. p is used to select *RAs* and *EFs* to form different prompts. The *Gaze tracking* state controls the *Tracking* function only, which represents the gaze-tracking module.

Two pure signals “Time out (TO)” and “Target hit” are used to change the prompts and terminate the LTM-RI trial. A pure signal either absents (no event), or presents (an event happens) any time $t \in \mathbb{R}$ [38]. If the gaze-tracking module detects gaze direction toward the target monitor within 7 s from the beginning of each prompt, a “Target hit” event is generated, which triggers the state transition to *Reward*. If no target hit is detected, TO event is generated. TO is combined with p to guide the transition in the AND substates of *Execution*. Once the state transition is done, p is increased by 1 to mark the next level of prompt. If prompt level 6 is completed without “Target hit,” the system transits to *Termination*.

IV. Experimental User Study

Fig. 6 shows the experiment room. The participant was seated in a wooden chair. The robot was placed in front of the participant, standing on a platform 32 cm above the floor. When the participant was seated, his/her eyes were approximately as high as the robot’s face. The two monitors and the robot were all 2 m away from the participant.

A. Participants

The NORRIS was tested by 14 children (12 males and 2 females; 12 Caucasians, 1 African American, and 1 Asian) with ASD. They were recruited from a research registry of the Vanderbilt Kennedy Center, and this study was approved by the Vanderbilt University Institutional Review Board. The participants were diagnosed with ASD by an expert clinician based on the DSM-5 [3] criteria. The diagnoses were done on average of 9.6 (SD: 5.9) months prior to enrollment in this study. Table II lists the statistical characteristics of the participants. The characteristics of individual participants are listed in the Appendix. Scores from gold standard diagnostic instruments (i.e., Autism Diagnostic Observation Schedule–Second Edition (ADOS-2) [27] and Mullen Scales of Early Learning [39]) were administered at that time. The participants met the spectrum cutoff on the ADOS-2. We use the early learning composite derived from the mullen scales of early learning to indicate existing intelligent quotient (marked as IQ in Table II). Parents of these children also completed the Social Responsiveness Scale–Second Edition [40] and Social Communication Questionnaire Lifetime Total Score (SCQ) [41] to index current ASD symptoms.

TABLE II
Statistical Characteristics of Participants

	ADOS-2 raw score	IQ	SCQ	SRS-2 T score	Age at enrollment (years)
Avg	21.29	54.71	14.86	63.36	2.78
SD	4.61	8.17	5.56	8.63	0.65

B. Experimental Procedure and Measurements

Four sessions were arranged on different days for each participant. On average, the time required to complete all four sessions was 27 days. Each session involved eight repeated LTM-RI trials as introduced in Section III-B. The left or the right monitor was randomly assigned as the target for each trial.

1) Preferential Attention: First, we evaluated the participants’ attention on the robot and the monitors, a measure which reflected their engagement. A region of interest was defined for each object that covered that object with a margin of 20 cm around it. We analyzed how their attention was distributed among the robot and the two monitors. We anticipated that participants would: 1) pay significant attention to the robot because the robot was the main interactive agent; and 2) pay more attention to the target monitor than the nontarget monitor, because the target monitor was referred to by the robot and displayed visual stimuli. We also tracked the change in the participants’ attention on the robot over the four sessions. We anticipated that if the participants’ interest in the robot sustained over the sessions, then their time spent looking at the robot would not change significantly.

2) Joint Attention Performance: Second, we evaluated the participants’ joint attention performance, which reflected the effectiveness of the system. For each session, we computed: 1) the number of trials in which the participants hit the target successfully; and 2) the average prompt levels the participants needed in order to hit the target. We anticipated that the participant’s performance would improve significantly if the robotic intervention was effective. In addition, we computed the intensity (defined in (2)) of each prompt level within the sessions. If the participants’ joint attention skill improved, we would see higher intensity values in low-prompt levels than in high-prompt levels.

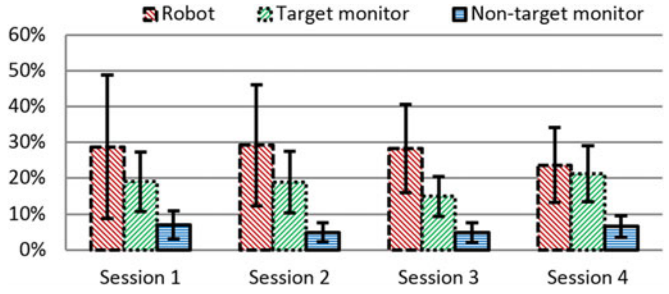


Fig. 7. Percentage of the session time participants spent looking at the robot, target monitor, and the nontarget monitor.

C. User Study Results

All 14 participants completed the four sessions, and thus the completion rate was 100%. This result is very promising when compared with other technology-assisted studies [22], [42]. We used the Wilcoxon-signed rank test for statistical analysis.

1) Preferential Attention: On average, in sessions 1–4,the participants spent 54.84%, 52.93%, 47.97%, and 51.58% of the session duration looking at the main objects (i.e., the robot, the target monitor, and the nontarget monitor), respectively. Fig. 7 shows the percentage of session durations that the participants spent looking at each of the main objects across the sessions. On average, in sessions 1–4, the participants spent 28.76%, 29.17%, 28.23%, and 23.69% of the session duration looking at the robot, respectively. In sessions 1–4, they spent 19.04%, 18.85%, 14.91%, and 21.31% of the session duration looking at the target monitor, respectively. We can see that the participants looked at the robot more than the monitors in every session. As expected, the participants looked at the target monitor much more than the nontarget monitor. The main reasons, we believe, were as follows: 1) the target monitor was referred to by the robot, and the participants responded more to the referred direction; 2) the target monitor displayed visual stimuli during prompts and rewards, which caught and held the participants' attention. Results showed that the participants spent very small portions of the session duration looking at the nontarget monitor (7.05%, 4.92%, 4.83%, and 6.58% in sessions 1–4, respectively).

We compared the time that the participants spent looking at the robot across all sessions, and found no statistically significant change ($p = 0.9515$–0.1937). This result suggests that the participants' interest in the robot held over the course of the sessions. The change in attention duration on the target monitor was also not statistically significant, except between sessions 2 and 3 ($p = 0.0203$), and between sessions 3 and 4 ($p = 0.0009$). In each session, different sets of static pictures, audio clips, and video clips were presented in the prompts. We noticed that the participants had different preferences for certain stimuli (e.g., one participant liked "Scooby Doo" more than "Dora"). Therefore, the fluctuation in attention time on the target monitor might be attributable to the change of stimuli. In addition, the attention time on the target monitor was statistically significantly higher than the nontarget monitor (p ranges from 0.0001 to 0.0006) in all four sessions.

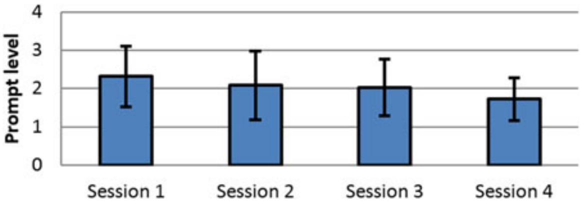


Fig. 8. Average target hit prompt levels in all sessions.

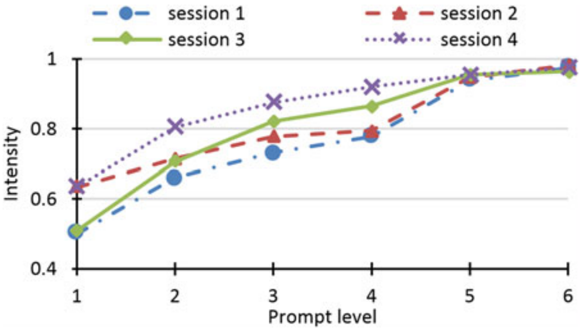


Fig. 9. Intensity of prompt levels across four sessions.

In summary, these results indicated: 1) among the three main objects, the participants paid most of their attention to the robot; 2) the participants' initial interest in the robot were maintained across the sessions; and 3) they paid significantly more attention to the target monitor than the nontarget monitor. Due to the heterogeneous neurodevelopmental trajectory and behavioral pattern of children with ASD, the participants had quite different attention patterns and joint attention capabilities. Therefore, Fig. 7 shows large standard deviations in all cases. The standard deviation of the looking time on the robot decreased from sessions 1 to 4, which showed that the participants' looking toward the robot tended to be stable after a few sessions' of intervention. However, this pattern was not shown for the two monitors.

2) Joint Attention Performance: Fig. 8 shows the average prompt levels participants needed to hit the target. Note that the lower the prompt level needed by participants, the better their performance was. We observe that from sessions 1 to 4, the average target hit prompt level decreased monotonically from 2.31 to 1.71. A Wilcoxon-signed rank test showed that the decrease in prompt level from sessions 1 to 4 was statistically significant ($p = 0.0115$). This indicated that the participants' performance improved significantly.

We further evaluated how the incremented LTM prompt levels elicited target hit behavior, i.e., whether the two goals discussed in Section III-A were achieved. The computation can be performed as follows:

$$P\left(ER_n | PT_{n-1}, \ldots PT_1\right) = \frac{number\,of\,trials\,ended\,with\,targethit\,on\,prompt\,leveln}{total\,number\,of\,trials}. \quad (6)$$

Then, the intensity of prompt level n can be computed according to (2). Here, we mark the intensity of prompt level n in session x as I_n^x. Fig. 9 shows the values of I in sessions 1–4.

We observe that $I_a^x > I_b^x$ (given $a > b$) in all four sessions. This indicated that the adding more instructive prompt levels on top of low prompt level elicited more target hits. Therefore, *Goal 1 was achieved***.** Fig. 9 also shows that for the same prompt level, $I_a^x > I_a^y$ or $I_a^x \approx I_a^y$ (given $x > y$) in most of the cases. The only exception is that I_1^2 is apparently higher than I_1^3. This means that, in general, as the participants had received more interventions in later sessions, their chances of a target hit on the same prompt level increased with occasional fluctuations.

I_6^1 to I_6^4 were 0.97, 0.98, 0.96, and 0.97, respectively. This leads us to conclude that the LTM-RI trial could eventually help participants hit the target in almost all of the trials. Thus, the prompt level content and the number of the levels (IN = 6) were properly designed. Therefore, *Goal 2 was also achieved*.

We also recorded subjective observation of these children's responses and feedback during the interaction with the robot. Due to the large heterogeneity of the autism spectrum, one child with ASD can be quite different from another. We found a spectrum of different behaviors from the participants. In most sessions, participants actively responded to the robot. Sometimes, they even tried to talk to the robot and imitate its gestures and speech. In some other sessions, participants quietly responded to the robot. There were also sessions where participants seemed tired and not quite interested in the interaction.

V. Conclusion and Discussion

In this paper, we introduced a new noninvasive autonomous robot-mediated joint attention intervention system, NORRIS. A humanoid robot was embedded as the intervention administrator. The looking behavior of the participants in response to robot prompts was detected using a new noncontact gaze-tracking method, which could track the participants' real-time gaze direction in a large range. The prompts were designed and implemented based on the presented LTM-RI prompting hierarchy, which is a general model of robot-mediated intervention for children with ASD.

NORRIS was validated through a four-session study. Fourteen children with ASD were recruited and all of them successfully participated in all the sessions. We measured their preferential attention toward the robot, target monitor, and the nontarget monitor. Results showed that the participants looked at the robot longer than other objects and this interest did not change significantly over the sessions. As expected, the participants paid significantly more attention toward the target monitor than the nontarget monitor in all the sessions. We also evaluated their joint attention performance. Results showed that the participants' performance improved significantly after the four intervention sessions. The results also proved the effectiveness of the LTM-RI model, i.e., the higher the prompt level, the higher the probability that target hit was achieved by the participants, and the participants could hit the target eventually in almost all the trials.

Therefore, we conclude that this study has two major contributions: 1) the design and development of a new joint attention intervention system with the introduction of the LTM-RI model; and 2) a user study with repeated observations which validated the effectiveness of NORRIS and LTM-RI.

However, it is important to notice that the current study also had limitations that need to be addressed in the future. First, NORRIS was only validated by a small group of children with ASD. In order to thoroughly evaluate the efficacy of NORRIS, it needs to be tested in formal clinical studies in the future. Second, while we did assess promising joint attention skills within the system, we did not systematically compare such improvements with other methods. Third, the current study did not investigate if such training can be generalized to other interactions, especially in human–human interactions. To address this issue, pre- and post-tests should be conducted to observe how joint attention behaviors in children with ASD generalize before and after using the robot-mediated intervention systems. A possible way to do the pre- and post-tests is video recording joint attention behaviors of children with ASD as they interact with other humans and coding these videos afterward. Fourth, the current study repeated a straightforward LTM-RI procedure in a limited number of sessions. However, using the same interaction content repeatedly will eventually cause the ceiling effect (e.g., participants hit the target on the first prompt in most of the trials) and/or loss of interest (e.g., participants feeling tired of doing the same intervention again and again) after a large number of sessions. Therefore, once these issues are detected, new interaction content under the same interaction protocol or a completely new interaction protocol need to be adapted by the system to update and reinforce the training procedure. Future robot-mediated works would benefit from exploring differences in combinations of approach in protocols, such as breaking down the cueing hierarchy that has nonverbal components only and then nonverbal and verbal components. Finally, LTM-RI was proposed as a general interaction model to implement robot-mediated intervention for children with ASD. Although the current study successfully validated LTM-RI for joint attention, we did not test it thoroughly in other types of training. Therefore, the eventual value of LTM-RI will need to be verified through other interventions.

Despite these limitations, this work is the first to our knowledge to design and empirically evaluate the usability and feasibility of a noninvasive fully autonomous robot-mediated joint attention intervention system over repeated observations. The preliminary results of this work are promising. Note that the NORRIS system architecture, components, and the LTM-RI protocol can be adapted to address other core deficits in ASD (e.g., social orienting, response to name). Thus, this work provides a framework of how to design and implement an effective robot-mediated intervention system in general. It is important to note that we do not propose this technology as a replacement for existing necessary comprehensive behavioral intervention and care for young children with ASD. Instead, this platform represents a meaningful step toward realistic deployment of technology capable of accelerating and priming a child for learning in key areas of deficits. While there is still much work to be done, in the future, we hope such systems will be deployed into schools and clinics for further validation and eventually be used as an effective and convenient tool for educators and psychologists.

Appendix

The characteristics of individual participants are shown in Table III as follows.

TABLE III
INDIVIDUAL CHARACTERISTICS OF PARTICIPANTS

Participant index	ADOS-2 raw score	IQ	SCQ	SRS-2 T score	Age at enrollment (years)	Gender	Race
1	26	49	18	67	2.66	M	White
2	22	67	19	75	2.82	M	African American
3	22	49	14	60	2.45	M	White
4	26	49	23	72	2.95	M	White
5	24	49	15	62	2.58	M	White
6	24	50	14	61	2.64	M	White
7	14	59	13	57	2.59	M	White
8	18	54	15	68	2.79	M	White
9	21	52	12	61	3.06	M	White
10	26	49	16	68	4.53	F	Asian
11	27	55	25	75	3.53	M	White
12	15	59	3	43	2.26	F	White
13	14	49	8	54	2.26	M	White
14	19	76	13	64	1.78	M	White

REFERENCES

[1] J.-J. Cabibihan, H. Javed, M. Ang Jr., and S. M. Aljunied, “Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism,” *Int. J. Soc. Robot.*, vol. 5, pp. 593–618, 2013.

[2] D. L. Christensen *et al.*, “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2012,” *Surveillance Summaries*, vol. 65, no. 3, pp. 1–23, Apr. 1, 2016.

[3] *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*, Amer. Psychiatric Assoc., St. Louis, MO, USA, 2013.

[4] B. Robins, K. Dautenhahn, and J. Dubowski, “Does appearance matter in the interaction of children with autism with a humanoid robot?” *Interaction Stud.*, vol. 7, pp. 509–542, 2006.

[5] E. S.-W. Kim, “Robots for social skills therapy in autism: Evidence and designs toward clinical utility,” Ph.D. dissertation, Dept. Computer Science, Yale Univ., New Haven, CT, USA, 2013.

[6] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, “Autonomous robot-mediated imitation learning for children with autism,” in *Proc. 2014 IEEE Int. Conf. Robot. Autom.*, 2014, pp. 2707–2712.

[7] Z. E. Warren *et al.*, “Can robotic interaction improve joint attention skills?” *J. Autism Develop. Disorders*, vol. 45, pp. 3726–3734, 2015.

[8] P. Pennisi *et al.*, “Autism and social robotics: A systematic review,” *Autism Res.*, vol. 9, pp. 165–183, 2015.

[9] H. Kozima, M. P. Michalowski, and C. Nakagawa, “Keepon,” *Int. J. Soc. Robot.*, vol. 1, pp. 3–18, 2009.

[10] E. S. Kim *et al.*, “Social robots as embedded reinforcers of social behavior in children with autism,” *J. Autism Develop. Disorders*, vol. 43, pp. 1038–1049, 2013.

[11] J. Greczek, E. Kaszubksi, A. Atrash, and M. J. Matarić, “Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders,” in *Proc. 23rd IEEE Int. Symp. Robot Human Interactive Commun. 2014*, Edinburgh, U.K., Aug. 2014, pp. 561–566.

[12] K. Dautenhahn *et al.*, “KASPAR—A minimally expressive humanoid robot for human–robot interaction research,” *Appl. Bionics Biomech.*, vol. 6, pp. 369–397, 2009.

[13] H. Kozima, C. Nakagawa, and Y. Yasuda, “Children–robot interaction: A pilot study in autism therapy,” *Prog. Brain Res.*, vol. 164, pp. 385–400, 2007.

[14] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, “Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism,” *IEEE Trans. Auton. Mental Develop.*, vol. 6, no. 3, pp. 183–199, Sep. 2014.

[15] B. Robins and K. Dautenhahn, “Developing play scenarios for tactile interaction with a humanoid robot: A case study exploration with children with autism,” in *Social Robotics*. New York, NY, USA: Springer, 2010, pp. 243–252.

[16] D. Feil-Seifer and M. J. Matarić, “Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders,” in *Proc. Conf. Exper. Robot.*, 2009, pp. 201–210.

[17] Z. E. Warren and W. L. Stone, “Best practices: Early diagnosis and psychological assessment,” in *Autism Spectrum Disorders*, D. Amaral, D. Geschwind, and G. Dawson, Eds. New York, NY, USA: Oxford Univ. Press, 2011, pp. 1271–1282.

[18] P. Mundy, J. Block, C. Delgado, Y. Pomares, A. V. Van Hecke, and M. V. Parlade, “Individual differences and the development of joint attention in infancy,” *Child Develop.*, vol. 78, pp. 938–954, 2007.

[19] P. Mundy, M. Sigman, and C. Kasari, “A longitudinal study of joint attention and language development in autistic children,” *J. Autism Develop. Disorders*, vol. 20, pp. 115–128, 1990.

[20] L. B. Adamson, R. Bakeman, D. F. Deckner, and M. Romski, “Joint engagement and the emergence of language in children with autism and Down syndrome,” *J. Autism Develop. Disorders*, vol. 39, pp. 84–96, 2009.

[21] M. Sigman *et al.*, “Continuity and change in the social competence of children with autism, Down syndrome, and developmental delays,” *Monographs Soc. Res. Child Develop.*, vol. 64, pp. 1–114, 1999.

[22] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar, “A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 2, pp. 289–299, Mar. 2013.

[23] Z. Zheng *et al.*, “Impact of robot-mediated interaction system on joint attention skills for children with autism,” in *Proc. 2013 IEEE Int. Conf. Rehabil. Robot.*, Seattle, WA, USA, 2013, pp. 1–8.

[24] S. R. Leekam, C. Nieto, S. J. Libby, L. Wing, and J. Gould, “Describing the sensory abnormalities of children and adults with autism,” *J. Autism Develop. Disorders*, vol. 37, pp. 894–910, 2007.

[25] A. Steinfeld, O. C. Jenkins, and B. Scassellati, “The Oz of Wizard: Simulating the human for interaction research,” in *Proc. 2009 4th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2009, pp. 101–107.

[26] M. E. Libby, J. S. Weiss, S. Bancroft, and W. H. Ahearn, “A comparison of most-to-least and least-to-most prompting on the acquisition of solitary play skills,” *Behavior Anal. Pract.*, vol. 1, pp. 37–43, 2008.

[27] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, *Autism Diagnostic Observation Schedule—ADOS-2*, 2nd ed. Torrance, CA, USA: Western Psychol. Serv., 2012.

[28] P. Mundy, A. Hogan, and P. Doelring, *A Preliminary Manual for the Abridged Early Social Communication Scales (ESCS)*. Coral Gables, FL, USA: Univ. Miami, 1996.

[29] S. M. Anzalone *et al.*, “How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot,” *Res. Autism Spectr. Disorders*, vol. 8, pp. 814–826, 2014.

[30] *Aldebaran Robotics*. 2018. [Online]. Available: https://www.ald.softbankrobotics.com/en

[31] Z. Zheng *et al.*, “Design of a computer-assisted system for teaching attentional skills to toddlers with ASD,” in *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being*. New York, NY, USA: Springer, 2015, pp. 721–730.

[32] X. Xiong and F. De la Torre, “Supervised descent method for solving nonlinear least squares problems in computer vision,” arXiv preprint arXiv:1405.0601, 2014.

[33] D. J. Feil-Seifer and M. J. Matarić, “A simon-says robot providing autonomous imitation feedback using graded cueing,” in *Proc. Int. Meeting Autism Res.*, 2012.

[34] B. Huskens, R. Verschuur, J. Gillesen, R. Didden, and E. Barakova, “Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention,” *Develop. Neurorehabil.*, vol. 16, pp. 345–356, 2013.

[35] Z. Zheng, E. Young, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, “Robot-mediated imitation skill training for children with autism,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 6, pp. 682–691, Jun. 2015.

[36] M.-G. Kim *et al.*, “Designing robot-assisted pivotal response training in game activity for children with autism,” in *Proc. 2014 IEEE Int. Conf. Syst., Man Cybern.*, 2014, pp. 1101–1106.

[37] D. Harel, “Statecharts: A visual formalism for complex systems,” *Sci. Comput. Program.*, vol. 8, pp. 231–274, 1987.

[38] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems: A Cyber-Physical Systems Approach.* Cambridge, MA, USA: MIT Press, 2011.

[39] E. M. Mullen, *Mullen Scales of Early Learning: AGS Edition*. Circle Pines, MN, USA: Amer. Guid. Serv., 1995.

[40] J. N. Constantino and C. P. Gruber, *The Social Responsiveness Scale*. Los Angeles, CA, USA: Western Psychol. Serv., 2002.

[41] M. Rutter, A. Bailey, and C. Lord, *The Social Communication Questionnaire*. Los Angeles, CA, USA: Western Psychol. Serv., 2010.

[42] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, “A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with autism,” *Int. J. Soc. Robot.*, vol. 6, pp. 45–65, 2014.

Zhi Zheng (S’09–M’16) received the Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2016.

She was a Research Assistant Professor of electrical engineering with Michigan Technological University, Houghton, MI, USA, from September 2016 to August 2017. Then, she joined the University of Wisconsin-Milwaukee, Milwaukee, WI, USA, as an Assistant Professor of biomedical engineering. Her research interests include human–machine interaction and human-centered computing.

Huan Zhao received the B.S. degree in automation from Xi’an Jiaotong University, Xi’an, China, in 2012, and the M.S. degree in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2016, where she is currently working toward the Ph.D. degree in electrical engineering.

From 2012 to 2014, she did research in the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University. Her research interest includes the design and development of systems for special needs using virtual reality, human–machine interaction and robotics.

Amy R. Swanson received the M.A. degree in social science from the University of Chicago, Chicago, IL, USA, in 2006.

She is currently a Research Analyst with Vanderbilt Kennedy Center’s Treatment and Research Institute for Autism Spectrum Disorders, Nashville, TN, USA.

Amy S. Weitlauf received the Ph.D. degree in psychology from Vanderbilt University, Nashville, TN, USA, in 2011.

She completed her predoctoral internship at the University of North Carolina, Chapel Hill. She then returned to Vanderbilt, first as a Postdoctoral Fellow and is currently an Assistant Professor of pediatrics with Vanderbilt University Medical Center, Nashville. She is also a Clinical Psychologist.

Zachary E. Warren received the Ph.D. degree in clinical psychology from the University of Miami, Miami, FL, USA, in 2005.

He is currently an Associate Professor of pediatrics and psychiatry with Vanderbilt University, Nashville, TN, USA. He is the Director of the Treatment and Research Institute for Autism Spectrum Disorders, Vanderbilt Kennedy Center, Nashville.

Nilanjan Sarkar (S’92–M’93–SM’04) received the Ph.D. degree in mechanical engineering and applied mechanics from the University of Pennsylvania, Philadelphia, PA, USA, in 1993.

In 2000, he joined Vanderbilt University, Nashville, TN, USA, where he is currently a Professor of mechanical engineering and electrical engineering and computer science. His current research interests include human–robot interaction, affective computing, dynamics, and control.

Dr. Sarkar is a Fellow of the ASME.