Effects of Responding to, Initiating and Ensuring Joint Attention in Human-Robot Interaction

Chien-Ming Huang and Andrea L. Thomaz

Abstract— **Inspired by the developmental timeline of joint attention in humans, we propose a conceptual model of joint attention with three parts: responding to joint attention, initiating joint attention, and ensuring joint attention. We conduct two experiments to investigate effects of joint attention in human-robot interaction. The first experiment explores the effects of *responding to* joint attention. We show that a robot responding to joint attention improves task performance and is perceived as more competent and socially interactive. The second experiment studies the importance of *ensuring* joint attention in human-robot interaction. We find that a robot's ensuring joint attention behavior is judged as having better performance in human-robot interactive tasks and is perceived as a natural behavior.**

I. INTRODUCTION

Joint attention is the process of sharing one's attention with another, using social cues (e.g., gaze). It is recognized as a crucial component in infant development. For example, it is thought that the failure to develop this fundamental social skill leads people on the autism spectrum to often have difficulties in communication and social interaction [1]. Thus, to facilitate natural human-robot interaction (HRI), we believe a basic social skill needed is the ability to respond to, initiate, and maintain joint attention with human partners.

This is a complex social skill for cognitive robots, and in our work we divide the skill into its three main components: responding to joint attention (RJA), initiating joint attention (IJA), and ensuring joint attention (EJA) to reflect psychological findings and behavioral observations [2], [12]. *Responding* is the ability to follow another's direction of gaze/gestures to attain a common perceptual experience. *Initiating* is the ability to manipulate another's attention in order to share an experience. *Ensuring* is the ability to monitor another's attention to verify that joint attention is reached and maintained. These correspond to developmental milestones. Infants start with the skill of following a caregiver's gaze, and then they exhibit pointing gestures to get the caregiver's attention. The initiating actions often come with an ensuring behavior, looking back and forth between the caregiver and the object [2]. Gorillas show similar monitoring and ensuring behaviors to check if an experimenter was attending to their actions and to solicit help [3].

We conducted two experiments to investigate different aspects of joint attention in HRI. The first experiment explores the effects of responding to joint attention. Results showed that a robot responding to joint attention is more transparent, such that interactive task performance is faster and more efficient, and people are more confident in the robot's task success. The second experiment studies the importance of ensuring joint attention. Results showed that ensuring joint attention is judged as yielding better task performance and is viewed as a natural behavior.

II. RELATED WORK

Most works in realizing joint attention focus on responding to joint attention. There are two main approaches to the problem. One is to build a constructive or learning model of developmental joint attention such that an agent learns the RJA skill through interactions [4], [5]. The other is to build a modular model of joint attention where the RJA skill is preprogrammed [6], [7].

Some work in realizing joint attention has also addressed aspects of initiating joint attention [8], [9]. These works implement IJA with preprogrammed mechanisms using eye gaze and pointing gestures to direct people's attention. To our knowledge, ours is the first work to explicitly look at ensuring joint attention and its role in facilitating HRI.

Prior work [10] probed effects of nonverbal communication in human-robot teamwork and suggested that implicit nonverbal communication positively impacts human-robot task performance. RJA involves nonverbal social cues, such as eye gaze, which acts as transparent communication. We present similar findings, and additionally test participants' confidence in task performance.

In a recent study on engagement, Rich et al. implemented a model for recognizing engagement in HRI [11], which has a significant overlap with joint attention in interaction. In particular, the event of directed gaze involves aspects of IJA and RJA. Mutual facial gaze concerns EJA, and adjacency pairs are acts that establish connections between interacting agents. Their work has focused on recognition instead of generation of these engagement behaviors.

III. JOINT ATTENTION IN HRI

In this section we operationalize the joint attention episode and describe our approach and implementation of this process on a robotic platform.

A. A Joint Attention Episode

A canonical joint attention episode, between two agents, can be described in five steps. First, two agents need to connect, to be aware of each other and to anticipate an upcoming interaction. The importance and the need of establishing a connection between interacting agents were pointed

Chien-Ming Huang (cmhuang@gatech.edu) and Andrea L. Thomaz (athomaz@cc.gatech.edu) are with School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.

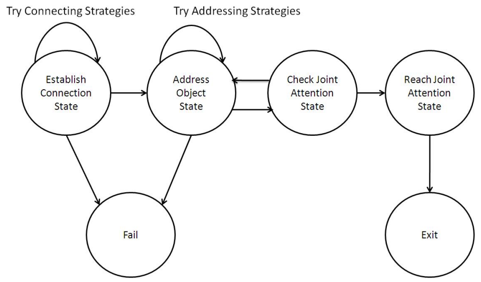
Try Connecting Strategies
Try Addressing Strategies
Establish Connection State
Address Object State
Check Joint Attention State
Reach Joint Attention State
Fail
Exit

Fig. 1. An integrative model of IJA and EJA for a joint-attention event.

out in [8], [13]. Second, the initiating agent makes a joint attention request by switching her attention to the object she intends to address. The initiating agent then addresses the object using communicative channels such as pointing gestures and/or voice. Third, the other agent responds to the request by attending to the referential focus. Fourth, after initiating joint attention, the initiating agent normally looks back and forth between the responding agent and the referential object to verify their response. This process is very fluid, the monitoring process is a quick glance, and the initiating agent may do the monitoring and addressing processes simultaneously (i.e., switching gaze while pointing to the focus). If the responding agent is not attending to the referential object, the initiating agent normally tries different strategies to get attention from the responding agent, including using bigger gestures or emphasizing gestures and making sounds. Finally, the two agents reach joint attention, both attending to the referential focus, and then continue the interaction. Importantly, the initiating agent does the ensuring joint attention process (step 4) periodically during the interaction to maintain joint attention.

B. *Our Approach*

Responding to joint attention involves knowing where the other agent's attention is directed. This is conveyed in many ways including eye gaze, head orientation, body pose, pointing gestures, or referential words. Normally, an agent uses a combination of several methods to draw attention from another agent. In our implementation, the RJA component is aware of pointing gestures and referential words.

An agent who intends to initiate a joint attention event should know the blueprint of the interaction she is going to start. In our implementation, an initiating agent follows a script that specifies actions that they intend to carry out, phrases to say and expectations from the responding agent, and joint-attention events. Each joint-attention event is executed by the finite-state-machine model, as shown in Fig. 1. A joint-attention event specifies the referential location and utterance for addressing the focus. Note that there may be several joint-attention events throughout an interaction.

To initiate joint attention, an agent starts with establishing a connection to the other agent. We implemented a set of addressing strategies including eye gaze, pointing gestures, and utterance. This is also seen in related work [8]. Langton argued that in addition to eye gaze, head orientation and pointing gestures are important cues to the direction of another's attention [14]. After addressing the focus, the agent ensures joint attention by first checking whether or not joint attention is reached. If not, the agent selects the next available addressing strategy until no strategies are available (i.e., ending in failure to reach joint attention).

Ensuring joint attention has two parts: *Monitoring* is the behavior of looking back and forth, checking the other agent's focus. *Soliciting* is using addressing strategies with increasing commitment to ensure joint attention is reached. Moreover, EJA can be categorized into two types based on when it occurs: 1) Initial EJA, which happens right after IJA to ensure its success, and 2) Periodical EJA, which happens throughout the interaction to ensure the other agent is still attending to the referential focus.

C. *Platform*

The robotic platform for this research is the Simon robot (Fig. 2), an upper-torso humanoid robot with two 7-DOF arms, two 4-DOF hands, and a socially expressive head. Simon has two 2-DOF eyes, eyelids and expressively colorful ears as another channel of communication. Simon can communicate attention by turning its head, eyes, and torso and can use its arms/hands for pointing gestures.

Simon has three primary ways to perceive the human partner. We use Microsoft's SAPI for speech recognition, with a small grammar designed for our experimental tasks. For pointing recognition, we made a paper pointer with a ARToolKit marker. Participants used this to point to objects. Additionally, Simon tracks a participant's face by keeping a detected face at the center of its eye camera view. Particularly, we used the face detection utility (Haar Cascade classifier) in OpenCV and applied criteria to filter out false positive recognitions. When idle, Simon does face tracking to stay engaged with participants.

IV. Effects of Responding to Joint Attention

Our first experiment looks at the impact that responding to joint attention has on a human-robot collaborative task.

A. *Hypotheses*

This experiment tests the following hypotheses:

- H1: People have a better mental model of a robot when it responds to joint attention requests.
- H2: People perceive a robot responding to joint attention as more competent.
- H3: People perceive a robot responding to joint attention as more socially interactive.

The first hypothesis tries to see if a robot responding to joint attention is more transparent to people, which should help people understanding the robot's intention. We use task-based metrics to show this improvement. The last two are about people's perception of RJA. We use questionnaire-based metrics to show this improvement.

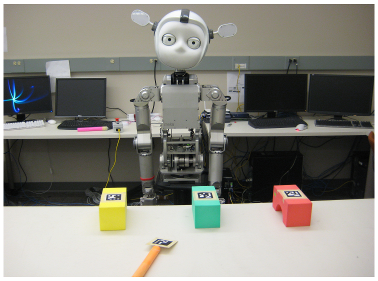

Fig. 2. The Simon robot in the RJA experimental setting.

B. Experimental design

Participants were given a labeling task to associate colors and names with objects (yellow for banana, green for watermelon, and red for apple). Participants sat across a table to interact with Simon and used a paper pointer and speech, as shown in Fig. 2. A white board nearby listed phrases used in the interaction, for participants' reference.

They completed the labeling task in two phases, first they labeled each object with its color (e.g., "This is a yellow object."). Then they gave each colored object a name (e.g., "The yellow object is a banana."). At any point during the interaction people were able to test Simon with questions (e.g., "Can you point to the yellow object?"). The two layers of labels makes the name concept depend on the color concept. Hence, if Simon has not learned the corresponding color concept, then it cannot learn the name concept. This design makes errors in the interaction more explicit.

C. Experimental conditions

To see how RJA affects performance in an interactive task and people's perception of the robot, we use a between-subject design with two groups:

- **With-RJA**: In this group, Simon responds to referential foci (i.e., a pointed to or a talked about known object) by gazing at it. If a referential concept has not yet been learned, Simon stays focused on the participant. When a participant tests an unknown concept, Simon gazes across all the objects.
- **Without-RJA**: In this group, instead of responding to referential foci, Simon stays focused on the participants as they teach the concepts.

In both groups, Simon has two basic behaviors. First, Simon always tracks a participant's face when not paying attention to a referential focus. Second, Simon's ears blink when hearing an utterance. The blinking is not only a way to tell a participant that the speech recognition engine is working but also to make the baseline behavior exhibit some interaction awareness. Note that ear blinking does not mean that Simon understands the concept or what a participant says, this was explicitly explained to participants.

D. Measures

We have four quantitative measures to evaluate our first hypothesis. If people have a good mental model of what the robot does and does not know, we expect to see a faster and more efficient teaching interaction in the following respect:

- M1: Number of errors during the teaching phase
- M2: Number of steps before recovering from errors
- M3: Number of redundant labels
- M4: Number of confirmations during the teaching phase

The interaction is simple enough that participants do not have trouble following the instructions. The underlying cause of any errors (M1, M2) is perceptual problems, either pointing recognition or speech recognition, and this manifests itself in two error cases: 1) when a participant requests a confirmation of a concept that has not been learned, indicating a failed labeling attempt previously; or 2) when they teach a concept different from the ground truth (i.e., labeling the yellow object as an apple), indicating that the wrong color label has been attached to the object.

A redundant label (M3) is not an error, but labeling an object more than once is an inefficiency in the teaching process. Presumably this is because the human is not sure whether or not the robot learned the label. Note that a label attempt is not logged as a redundant if it is a repetition due to a speech recognition error. A confirmation (M4) is asking the robot to explicitly point to a particular object.

We expect the transparency of RJA to improve people's ability to detect and recover from errors (reducing M1 and M2), and to improve people's understanding of the robots internal state (reducing M3 and M4).

After the interaction, participants completed a 7-question survey (see Fig. 3). Afterwards they answered a survey with open-ended questions. In addition to these specific metrics, we also collected video of the interactions, we use this to give some anecdotal insight into our findings.

E. Results

We had Twenty-four participants. Four were discarded due to either speech recognition engine, vision software, or control software failures during the interaction. All of the valid 20 participants were students from the local campus population and were randomly assigned to groups (10 in each). We use the Student's t-test in our data analysis.

Table 1 summarizes results of the quantitative measures, all of which were significantly different between the groups. The significant difference on total number of errors (M1) was mainly due to participants in the without-RJA group lacking a good mental model of the robot, usually resulting in teaching the robot too fast. Moreover, the result of M2 showed that it took longer for participants in the without-RJA group to identify and correct errors. This evidence suggests that RJA serves as a good transparency device to help participants understand the robot. These results about error detection and recovery also confirm prior work on nonverbal interaction with the Leonardo robot, where people went too fast, and eye gaze was beneficial in helping people notice errors early and correct them [10].

Getting joint attention responses from Simon helped participants in the with-RJA group understand whether Simon had learned the concepts or not. This was supported by our

TABLE I

RJA experiment, quantitative results. All measures are significantly reduced in the with-RJA group.

	with-RJA n=10		without-RJA n=10		Significant level	
	Mean	S.D.	Mean	S.D.	t	p<
M1	0.2	0.42	2.9	2.51	4.98	0.001
M2	4	2.83	22.56	35.47	10.27	0.001
M3	2.8	3.74	7.8	10.69	6.00	0.001
M4	4.5	1.18	9.8	5.98	6.27	0.001

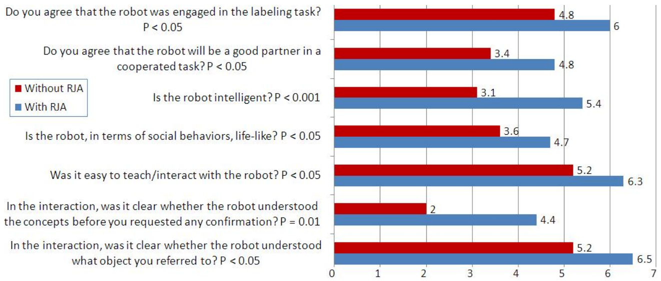

Fig. 3. Results of the post-experiment questionnaire for the RJA experiment.

findings with M3, that participants in the without-RJA group had significantly more redundant labels than the with-RJA group. Lack of responses from Simon caused participants to label multiple times to ensure that Simon learned the concepts. Similarly, participants in the without-RJA group requested more confirmations from Simon until they felt confident that Simon learned the concepts (M4). In contrast, in the with-RJA group, participants requested confirmations less than six times on average (six would be needed to do one for each concept). This showed that with-RJA participants understood of the robot's internal state during the process.

In addition to the quantitative measures, results of the post-experiment questionnaire (Fig. 3) shed light on how participants perceived the interaction. The with-RJA group was significantly more positive to all questions than the without-RJA group. Specifically, results from the last two questions "Was it clear whether or not the robot understood the concept before you requested confirmations?" and "Was it clear whether or not the robot understood which object you referred to?" both upheld H1, about improved mental models. Results from the questions about ease of interaction with the robot, life-like social behavior, perception of intelligence, and potential for being a good partner in collaborative settings supported H2. Results of questions about ease of interaction, potential of being a good collaborative partner, and engagement in the label task supported H3.

Moreover, results from a self-report survey were also consistent with quantitative measures and the questionnaire results. Most participants in the with-RJA group mentioned that Simon used head/eye movements to convey attention/awareness, while participants in the without-RJA group commented that Simon is socially strange. Most participants in the with-RJA group noted that gaze or head orientation were the cues they used to verify if Simon had learned the concepts, whereas without-RJA participants said that they were frustrated, had a hard time telling whether Simon had learned, and/or felt they were being ignored.

Looking back at the video collected during the interactions, we find several behaviors that give additional insight into joint attention for HRI. Two observations were common across both groups. First, participants looked back and forth between the referred object and Simon's face to see if Simon understood the concepts—ensuring joint attention. This leads in to our next experiment and the hypothesis that EJA is needed in natural HRI. Second, participants showed RJA themselves when Simon initiated a joint attention event (i.e., they followed Simon's pointing gesture).

V. The importance of Ensuring Joint Attention

In our second experiment, we look at the importance of initiating and ensuring joint attention (IJA/EJA) for HRI.

A. Experimental Design

As described in Sec. III, we break EJA into two components: *monitoring* the gaze of the interaction partner, and *soliciting* the interaction partner's attention with various communication strategies. Additionally, once joint attention is established, periodical EJA is *periodically* checking that it is maintained. Our implementation of soliciting is a combination of eye gaze and a pointing gestures. In this experiment we systematically analyze the effect that monitoring, soliciting, and periodical EJA have on an HRI scenario. In a video-based experiment, participants are asked to rank collections of videos in which Simon used some or all of the three EJA aspects. Our video-based design controls the experiment such that we can focus on studying how people perceive ensuring joint attention instead of on the technical challenge of perceiving participants' attention. This experiment is intended to establish the importance of EJA in HRI for future research on building an autonomous EJA for interactive agents.

We hypothesize that ensuring joint attention affects task performance. Psychological findings [2], [12] and our observations in the previous experiment indicate that ensuring joint attention is a natural behavior. Therefore, we have two hypotheses:

- H4: A robot that ensures joint attention will be judged as having better task performance.
- H5: Ensuring joint attention is perceived as a natural behavior in social interaction with a robot.

B. Robot Scenario Manipulations

To test these hypotheses we use three service robot scenarios (Fig. 4). Each scenario is used to investigate different aspects of EJA: Soliciting, Monitoring, and Periodically verifying joint attention.

Presentation: The first scenario looks at both the initial and periodical ensuring joint attention event. In this scenario (Fig. 4(a)), Simon is a tour guide robot giving a presentation. Simon stands beside a poster and faces a person, greets

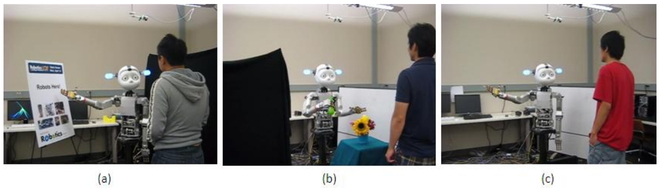

Fig. 4. EJA Scenarios. (a) presentation scenario, (b) reception scenario, (c) giving-directions scenario.

the person, and then gives a presentation about Robotics at Georgia Tech. When Simon is about to start the presentation, the person's cell phone rings, and the person walks away to take the call. Once finished, the person walks back to re-engage in the presentation.

In the video manipulations, Simon exhibited varying degrees of EJA behavior in the initiation and maintenance of joint attention in response to the cell phone distraction. We have four behavioral variations of this scenario:

V1: Monitoring + Soliciting + Periodical: This is the full EJA behavior. Simon switches gaze between the human and the poster to monitor joint attention, and waits until the person finishes their call to start the presentation. In addition, Simon periodically looks back to the human to ensure attention during the presentation.

V2: Monitoring + Soliciting + not Periodical: The robot ensures the initial joint attention (i.e., waits until the person comes back) but does not periodically check joint attention during the presentation.

V3: Monitoring + not Soliciting + Periodical: Simon ignores the person's phone call situation and continues the presentation. However, Simon periodically looks toward the person during the presentation.

V4: not Monitoring + not Soliciting + not Periodical: Simon continues the presentation when the person leaves and does not glance to them while presenting.

We expect the full EJA behavior is the most desirable in terms of both task performance and naturalness.

Receptionist: The second scenario focuses on monitoring and soliciting joint attention. In this scenario (Fig. 4(b)), Simon is a service robot receiving a guest at a reception desk. The robot has a secondary task of watering plants. A guest comes to the reception desk and asks for his friend Bob. The robot has the guest wait, and turns to deliver Bob the message. Bob is focused on work at his computer when the robot comes to him. The robot gives a prompt "Excuse me, sir" to get Bob's attention (i.e., tries to establish a connection). Bob hears and turns to the robot, but then accidentally drops a cup of coffee while turning around. He is distracted and tries to clean up the mess before continuing the interaction with the robot. If the robot does not ensure joint attention, it would deliver the message no matter if Bob is listening or not. If the robot ensures Bob has joint attention and actually received the message, it will be more effective.

We test three variations of the reception scenario:

V1: Monitoring + Soliciting: This is the full behavior. Simon monitors Bob's attention, notices Bob is distracted, and waits until Bob finishes cleaning to deliver the message.

V2: Monitoring + not Soliciting: Simon monitors Bob's attention but ignores his situation, and delivers the message when Bob is not paying attention.

V3: not Monitoring + not Soliciting: Simon turns to Bob, delivers the message, and goes back to the side task.

The fourth variation, soliciting but not monitoring, is infeasible in reality. We expect people to prefer the full behavior over the others with respect to task performance.

Giving directions: The third scenario focuses on periodical EJA. In this scenario (Fig. 4(c)), Simon is a guide robot directing a person to the restroom. A person comes to Simon and asks where the restroom is. Simon answers with directional speech and a directional gesture.

In the video manipulations, Simon always does the initial ensuring joint attention event (i.e., monitoring and soliciting). We have two behavioral variations in this scenario:

V1: Periodical: Simon looks back and forth between the person and the direction of the restroom while giving directions.

V2: not Periodical: Simon only looks toward the direction of the restroom throughout.

We expect that periodical EJA is more desirable in terms of naturalness.

C. Experimental Procedure

We want to see if varying degrees of EJA behaviors, play into people's perception of the robot and the interaction. Therefore, we use a within-subjects design to measure how people perceive the effectiveness of communication and naturalness of behavior.

Fifteen participants from the local campus population were recruited for this experiment. Participants watched three sets of videos, one for each scenario. We randomized the ordering of the behavioral variations within each scenario. To minimize order effects, we randomly sorted the videos into three groups (i.e., different orders of videos within scenarios), and people were randomly assigned to one of these three video groupings (five participants in each group).

Participants were directed to a website containing the first set of videos. They were told to watch each video all the way through the first time, since there were only slight differences between videos. Then they were allowed to watch the videos as many times as they wanted.

After watching all the videos for a particular scenario, participants filled out a survey regarding that scenario, comparing the variations. There were six questions for the presentation scenario and four questions for the reception and directions scenarios (detailed in the Sec. V-D).

D. Results

For each question, participants were asked to rank the videos in the scenario with respect to some quality, and we analyze people's "first choice" selection. We use a chi-square test for goodness of fit to determine if the distribution of first choice selections is significantly different from uniform.

1) EJA judged as better task performance: For both the presentation and the reception scenarios, participants were asked to rank videos based on the following task-performance qualities: (1) "How well the person in the videos can recall or receive the information from the robot." In the presentation scenario, the videos were not equally preferred ($\chi^2(3, 15) = 24.73, p < .01$), and similarly for the reception scenario ($\chi^2(2, 15) = 30, p < .01$). In both scenarios the full EJA variation was most desirable. (2) "How good the robot was at communicating information." Again the choices were not equally preferred in the presentation ($\chi^2(3, 15) = 30.6, p < .01$) or reception scenarios ($\chi^2(2, 15) = 30, p < .01$), and the full EJA was most desirable. This result supports H4, that a ensuring joint attention is judged to have better interactive task performance.

2) EJA judged to be more engaged: For the presentation and directions scenarios, participants were asked to rate "How well Simon engaged the person in the videos". The result from the presentation ($\chi^2(3, 15) = 37.53, p < .01$) and the directions scenario ($\chi^2(1, 15) = 8.07, p < .01$) showed that the videos were not equally preferred, and that the full EJA behavior is the most desirable.

Additionally, participants ranked the videos according to "How similar the robot's behaviors are to theirs if they were asked to perform the same task". The result from the presentation scenario ($\chi^2(3, 15) = 30.6, p < .01$) and the directions scenario ($\chi^2(1, 15) = 11.27, p < .01$) revealed that the full EJA behavior is the most similar behavior to theirs. Both of these results support H5 that ensuring joint attention is judged as a natural social behavior.

3) EJA generally preferred: For all scenarios, participants ranked the videos according to their preference "...if they were asked to design behaviors for a robot in similar scenarios". For the presentation ($\chi^2(3, 15) = 37.53, p < .01$), reception ($\chi^2(2, 15) = 24.4, p < .01$), and directions ($\chi^2(1, 15) = 8.07, p < .01$) scenarios videos were not equally preferred: 14 participants selected full EJA behavior as the most desirable for presentation and reception, while 13 participants choose EJA in the directions scenario. Thus people generally desire EJA behaviors on a robot.

In the survey about each scenario, participants were asked to comment on the differences they observed and how they liked/disliked the videos. These comments give us insight that it was in fact our manipulation of EJA behaviors that were playing into people's choices. For the presentation scenario, all participants commented about Simon making sure the person was paying attention before the presentation versus not. Twelve participants noted that Simon looked at the user occasionally versus not. Participants often used phrases like "make eye contact", "engage user", and "recapture attention" to describe the behavior. Similarly, most participants mentioned the two main differences in the reception scenario. And in the directions scenario, most participants (13) noticed the difference was whether or not the robot turned to the person during interaction, and many made positive comments about this. For example, "good communication" and "is mostly how normal people would behave." However, one participant described the behavior as "unnecessary head turns" showing an alternative perspective.

VI. Discussion

Our two experiments investigate different aspects of joint attention in HRI. We found that responding to joint attention helps people to understand the robot's internal state leading to fewer confirmations, fewer errors, fewer redundant labels, and faster error recovery during a teaching interaction. On the initiating side, people believe that ensuring behaviors improve communication which improves task performance. Additionally, people indicate that EJA behaviors as natural behaviors and that robots should have EJA behaviors to facilitate human-robot interaction. In this section we layout some of the challenges for joint attention in HRI.

A. Challenges in Perceiving Joint Attention

There are several technical challenges in realizing joint attention in the context of human-robot interaction, particularly, perceiving the human's referential focus. Eye gaze tracking is still limited and unreliable in open-world settings. In our work, we tried using an off the shelf eye-tracking solution, but found it too restrictive since a person has to limit her movement which greatly limits interactive scenarios. An alternative is to estimate eye gaze using head pose tracking. There have been research efforts tracking humans' head orientation in real-time [15]. We plan to use head orientation in estimating attention in future work.

Deictic gesture recognition is another challenge for joint attention. Humans use a variety of hand gestures to direct attention and facilitate interaction. Particularly pointing gestures are useful when referring to an object. In our experiment people used a paper marker as their pointing device, but clearly more naturalistic deictic gesture recognition is necessary for open-ended interaction.

B. Open Questions in our Approach

There are some issues that we have not addressed in our current approach. First, when and how frequently should a robot do ensuring joint attention in an interaction? Even though our second study suggests that EJA behavior is natural, we believe this only when it happens at a time and frequency that meet people's expectation. In our study this was given by our scenario design, but generating appropriate timing autonomously is an open question. Second, the model should handle interactions with a group of people. For example, instead of ensuring everyone in the group is paying attention, a robot may just need to engage most people in the group. In addition, the strategies for getting attention from a group may be different. Third, a robot should be able to learn strategies through interactions with humans and use strategies adaptively according to situations and the person it is interacting with.

C. Effects of Embodiment

The benefit of using an embodied platform for evaluation of a computational model of joint attention has been recognized. An embodied platform provides the capability of being physically interactive and is more likely to draw natural responses from participants. Moreover, in contrast to empirical observations, embodiment allows experiments to be repeatable, and different aspects are easily separated for evaluation [16].

However, research questions remain about how much the fact that Simon has a human-like head and eyes influences people's expectations. For example, in the responding to joint attention study, would people still get as much out of the response if it was not anthropomorphic? Also, it is unknown how much the fact that Simon use a human-like channel to convey attention affects people's understanding. Would people still perceive Simon the same way if it uses a not human-like channel, such as flashing its ears, to respond to joint attention? Future work is needed to investigate the effect of embodiment on our hypotheses.

VII. Conclusion

In this work, we use a conceptual model of joint attention consisting of responding to joint attention, initiating joint attention, and ensuring joint attention and have implemented aspects of this model on a robotic platform. We evaluated the effects of responding to joint attention, and found that robots responding to joint attention produce better task performance and are seen as more competent and socially interactive. We evaluated the importance of ensuring joint attention in HRI, and learned that robots ensuring joint attention are judged as having better performance in human-robot interactive tasks and are perceived as more natural.

References

[1] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*. Cambridge, Massachusetts: The MIT Press, 1997.

[2] P. Mundy and L. Newell, "Attention, Joint Attention, and Social Cognition," *Current Directions Psychological Science*, vol. 16, no. 5, pp. 269–274, 2007.

[3] J. C. Gomez, "The emergence of intentional communication as a problem-solving strategy in the gorilla," in *Language and Intelligence in Monkeys and Apes: Comparative developmental approaches*. Cambridge University Press, 1990, pp. 333–355.

[4] E. Carlson and J. Triesch, "A computational model of the emergence of gaze following," in *In Connectionist models of cognition and perception II*, H. Bowman and C. Labiouse, Eds. World Scientific, 2003, pp. 105–114.

[5] Y. Nagai, K. H. Y, A. Morita, and M. A. Y, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, pp. 211–229, 2003.

[6] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *In Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, 2001.

[7] A. L. Thomaz, M. Berlin, and C. Breazeal, "An embodied computational model of social referencing," in *In IEEE International Workshop on Human Robot Interaction*, 2005.

[8] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," *IEEE Transaction on Industrial Electronics*, vol. 50, no. 4, pp. 636–643, 2003.

[9] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*. Springer Berlin, 1999, pp. 176–195.

[10] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2005, pp. 383–388.

[11] C. Rich, B. Ponsleur, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*. New York, NY, USA: ACM, 2010, pp. 375–382.

[12] J. H. Williams, G. D. Waiter, O. Perra, D. I. Perrett, and A. Whiten, "An fMRI study of joint attention experience," *NeuroImage*, vol. 25, no. 1, pp. 133–140, 2005.

[13] T. Striano, V. M. Reid, and S. Hoehl, "Neural mechanisms of joint attention in infancy," *The European journal of neuroscience*, vol. 23, no. 10, pp. 2819–23, 2006.

[14] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? Cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50–59, 2000.

[15] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell, "Fast stereo-based head tracking for interactive environment," in *Int. Conference on Automatic Face and Gesture Recognition*, 2002.

[16] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Study*, vol. 7, no. 2, pp. 135–169, 2006.