Joint Attention in Human-Robot Interaction

Chien-Ming Huang and Andrea L. Thomaz
801 Atlantic Dr., Atlanta GA, 30332

Abstract

We propose a computational model of joint attention consisting of three parts: responding to joint attention, initiating joint attention, and ensuring joint attention. This model is supported by psychological findings and matches the developmental timeline in humans. We present two experiments that test this model and investigate joint attention in human-robot interaction. The first experiment explored the effects of responding to joint attention on human-robot interaction. We show that robots responding to joint attention are more transparent to humans and are more competent and socially interactive. The second experiment studied the importance of ensuring joint attention in human-robot interaction. Data upheld our hypotheses that a robot's ensuring joint attention behavior yields better performance in human-robot interactive tasks and ensuring joint attention behaviors are perceived as natural behaviors.

Joint attention, a process to share one's current attention with another by using social cues such as gaze, has been recognized as a crucial component in interactions and an important milestone in infant development. One of the key components of Autism Spectrum Disorder is the failure to develop joint attention capabilities. It is hypothesized that the failure to develop this fundamental social skill leads people on the autism spectrum to often have difficulties in communication and interaction with other people (Baron-Cohen 1997). Therefore, to facilitate natural human-robot interaction (HRI), we believe that robots need social skills to respond to, initiate, and maintain joint attention with humans.

Our model of joint attention reflects the complexity of this social skill for cognitive robots, dividing the skill into its three main components: responding to joint attention (RJA), initiating joint attention (IJA), and ensuring joint attention (EJA). RJA is the ability to follow another's direction of gaze and gestures in order to attain common experience. IJA is the ability to manipulate another's attention to a focus of interest in order to share experience. EJA is the ability to monitor another's attention and to ensure that the state of joint attention is reached. These match the developmental milestones of joint attention in infancy. Infants start with the skill of following a care-giver's gaze, and then they exhibit pointing gestures (especially declarative pointing gestures) to get the care-giver's attention. Importantly, the initiating actions often come with an ensuring behavior that is to look back and forth between the care-giver and the referential object (Mundy and Newell 2007)

We conducted two experiments to test our model and to investigate joint attention in HRI. The first experiment explored the effects of responding to joint attention. Results showed that a robot responding to joint attention is more transparent to people such that interactive task performance is better and people are more confident in the robot's task success. In addition, people perceive a robot responding to joint attention as more competent and socially interactive. The second experiment studied the importance of ensuring joint attention in HRI. Results suggested that ensuring joint attention yields better task performance and is viewed as a natural behavior in interactions.

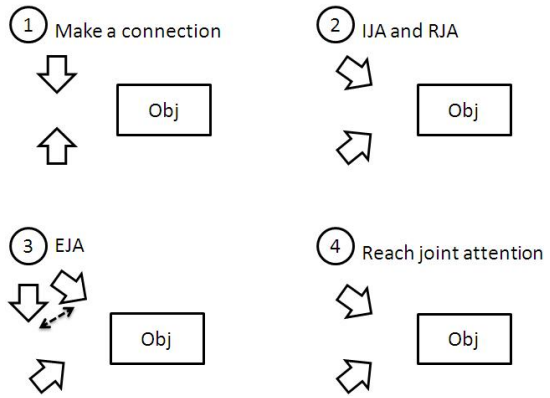
Related Work

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 (Kaplan and Hafner 2006). Additionally, embodiment provides access to internal states as a behavior develops (Deák, Fasel, and Movellan 2001).

Most works in realizing joint attention focused on aspects of responding to joint attention (RJA) only. They have approached the problem of RJA from two different angles. One is to build a constructive or learning model of developmental joint attention such that an agent learns the RJA skill through interactions (Carlson and Triesch 2003; Nagai et al. 2003). The other is to build a modular model of joint attention where the RJA skill is preprogrammed for an agent (Kozima and Yano 2001; Thomaz, Berlin, and Breazeal 2005).

Some work in realizing joint attention (Imai, Ono, and Ishiguro 2003; Scassellati 1999) has also addressed aspects of initiating joint attention (IJA). These works implement IJA with preprogrammed mechanisms involving eye gaze and pointing gestures. To the best of our knowledge, our

Figure 1: Joint attention in interaction. An arrow represents an agent and direction of an arrow indicates the agent's attentional focus.



work is the first to explicitly address aspects of ensuring joint attention and the role it plays in facilitating HRI.

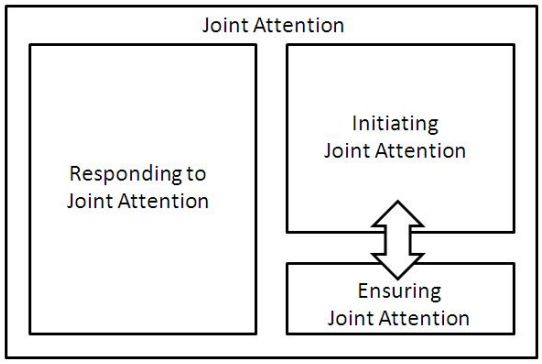
A similar work (Breazeal et al. 2005) probed effects of nonverbal communication in human-robot teamwork and suggested that implicit nonverbal communication positively impacts human-robot task performance. RJA involves nonverbal social cues, such as eye gaze, which acts as transparent communication. One main difference between their work and the first experiment presented here is that we use additional measures to test participants' confidence in task performance.

In a recent study on engagement, Rich et al. proposed and implemented a model for recognizing engagement in HRI (Rich et al. 2010). The concepts of engagement have 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.

Realization of Joint Attention

Figure 1 shows a simple depiction of joint attention. The interaction can be described as four steps. First, two agents need to connect to each other before an interaction. The purpose of this is to be aware of each other and to anticipate an upcoming interaction. Second, the initiating agent (the upper one in Fig. 1) initiates joint attention by switching her attention (i.e., gaze) to the object of interest. The initiating agent addresses the object using communicative channels such as pointing gestures or vocal comments. In the meantime, the other agent responds to this joint attention request by switching attention to the referential focus. Third, after initiating joint attention, the initiating agent normally looks back and forth between the responding agent and the referential object to ensure the responding agent attends to the joint attention. The initiating agent may do the ensuring and the 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 tries different ways (e.g., emphasizing gesture or making sounds) to get attention from the responding agent. Finally, the two agents reach joint attention while both attending to the referential focus and then continue the interaction. Importantly, the initiating agent does the ensuring joint attention process (step 3 and 4) periodically during the interaction to maintain joint attention.

Figure 2: A high-level structure of joint attention model.



A computational model of joint attention

We propose a joint-attention model consisting of three components: responding to joint attention (RJA), initiating joint attention (IJA), and ensuring joint attention (EJA) to reflect psychological findings and behavioral observations (Mundy and Newell 2007; Williams et al. 2005). Figure 2 shows the high-level structure of our model. In the model, RJA and IJA run exclusively. However, EJA is an always-on process interacting with IJA to ensure that the other agent attends to the right focus.

Joint attention involves an agent gazing at or turning to the object referred by the other agent. To do so, this agent first needs to know how the other agent conveys attention and to know where the other agent's attention is. An agent conveys attention using different methods including eye gaze, head orientation, body pose, pointing gestures, or referential words. Moreover, it is normal that an agent uses a combination of several methods at a time to draw attention from the other agent. In our current design, we assumed a responding agent knows the ways that an initiating agent conveys attention. In particular, we have implemented the RJA component to be aware of pointing gestures and referential words.

An agent who intends to initiate a joint attention should know the blueprint of the interaction she is going to start. In our design, an initiating agent follows a script that specifies actions that she intends to carry out, phrases she wants to say and she expects from the responding agent, and joint-attention events. A joint-attention event is executed by a finite-state-machine model as Figure 3 shown.

To initiate joint attention, an agent starts with establishing a connection to the other agent. The importance and the need of establishing a connection between interacting agents were pointed out in (Imai, Ono, and Ishiguro 2003;

Figure 3: An integrative model of IJA and EJA.

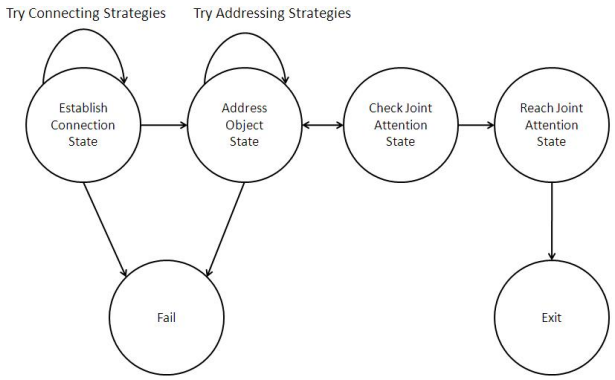
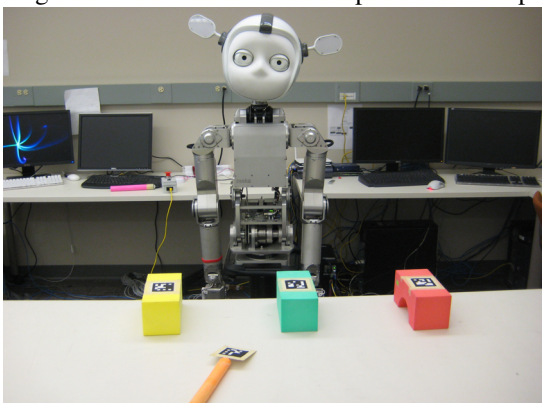


Figure 4: Simon and the RJA experimental setup.



Striano, Reid, and Hoehl 2006). A set of connecting strategies are designed to ensure a connection is established before further interaction. Connecting strategies include uttering (e.g., 'Excuse me') and using bigger gestures (e.g., waving).

Once a connection is established, the agent addresses the focus with communicative channels. The goal is to orient the other agent's attention to the referential focus so that the two agents could reach joint attention. We designed a set of communicative channels (i.e., addressing strategies) for relocating the other agent's attention. Addressing strategies include eye gaze, pointing gestures, and utterance. This design is supported by psychology and cognitive science findings, and peer design (Imai, Ono, and Ishiguro 2003). Langton argued that not just eye gaze is a central cue to the direction of another's attention, head orientation and pointing gestures are equivalently important (Langton, Watt, and Bruce 2000). Moreover, joint attention is supported by perception of multi-modal social cues, and infants respond more to pointing gestures than gaze (Deák, Fasel, and Movellan 2001). After addressing the focus, the agent checks 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). Otherwise, the agent continues the interaction.

Conceptually, EJA could be viewed as two parts: monitoring and ensuring. In practice, monitoring is the behavior of looking back and forth between the other agent and the referential object, and ensuring is using addressing strategies to make sure joint attention is reached.

Effects of Responding to Joint Attention on Human-Robot Interaction

We designed our RJA experiment to test the following three hypotheses (H1-H3):

- 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 more competent.
- H3: People perceive a robot responding to joint attention more socially interactive.

The first hypothesis tries to see if a robot responding to joint attention is more transparent to people. Transparency helps people's understanding of a robot's intention and capacity that further facilitate better HRI. The last two try to reveal that people perceive a robot responding to joint attention in a more positive way, which is important for improving the human-robot relationship.

Robotic platform

The robotic platform for this research is the Simon robot (Figure 4). Simon is an upper-torso humanoid robot with two 7-DOF arms, two 4-DOF hands, and a socially expressive head. In addition, Simon has two 2-DOF eyes, eyelids and expressively colorful ears as another channel for communication. Simon is capable of turning its head, eyes, and torso as paying attention and showing pointing gestures. We use Microsoft's SAPI for speech recognition, and use a small grammar designed for the experiment interaction to facilitate recognition.

Experimental design

Participants were given a labeling task to associate colors and names with objects (i.e., yellow for banana, green for watermelon, and red for apple). The interaction space was organized as shown in Figure 4. Participants sat across a table to interact with Simon and used pointing gestures (a paper pointer) and speech (e.g., "This is a yellow object.", "The yellow object is a banana.", and "Can you point to the banana?") to label and to test objects. Beside Simon was a white board that listed phrases used in the interaction for participants' reference.

Experimental conditions

We want to compare a robot with RJA to a robot without RJA and to see how RJA affects performance of an interactive task and people's perception of the robot. We use a between-subject design to compare two groups: with-RJA group and without-RJA group.

In the with-RJA group, Simon responds to referential foci (i.e., a pointed or a mentioned known object) by gazing at it. If a referential concept has not been learned yet, Simon will

stay focused on the participant. When a participant requests a concept that has not been learned, Simon gazes over all the objects. These gaze behaviors are the basic RJA mechanism for Simon to communicate with participants implicitly. In the without-RJA group, Simon, instead of responding to referential foci, 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 some utterance. The blinking is not only a way to tell a participant that the speech recognition engine is working but also to make Simon more life-like in terms of social awareness. Note that ear blinking does not mean that Simon understands the concept or what a participant says, and this was explicitly explained to experimental participants.

Measures

We had four quantitative measures (M1-M4) to evaluate H1.

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

An error is defined as when the human participants either requests a confirmation of a concept that has not been learned or teaches a concept different from the ground true (i.e., labeling the yellow object as an apple). A redundant label is when a participant has made the same label attempt before (no matter the concept had been learned correctly or not). Note that a label attempt did not count as a redundant if it was a repetition due to an utterance not being detected by the speech recognizer.

Results

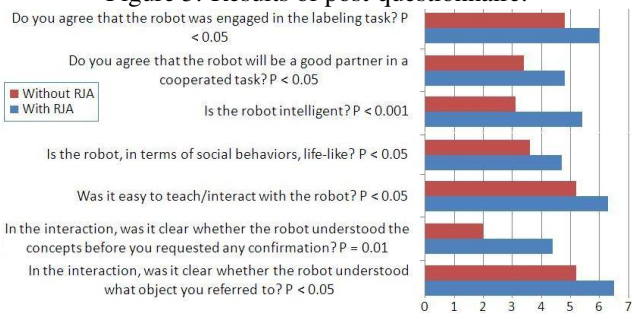
Twenty-four participants were recruited for this experiment. Four of them were discarded due to either speech recognition engine, vision software, or control software crashed during the interaction. All the valid 20 participants (19 males and 1 female) were students from the local campus population and were randomly assigned to either the with-RJA or the without-RJA group (10 in each group). A total of seven participants (four from the with-RJA group and three from the without-RJA group) reported that they did not have any experience related to robotics.

Table 1 summarizes results of the quantitative measures, all of which were significantly different between the groups. The significant difference on M1 was mainly due to participants in the without-RJA group tending to teach the robot at their own pace, which was usually too fast. Therefore, more errors were generated during the interaction. Moreover, the result on M4 showed that it took longer for participants in the without-RJA group to identify and to correct errors. This evidence suggested that RJA serves as a good transparency for participants to understand the robot. Without RJA on the robot, participants had hard time to build and to maintain a mental model of the robot. The results also confirm our prior work on nonverbal study with Leo, where people went too fast, and eye gaze was good for getting people to notice errors early and correct them (Breazeal et al. 2005).

Table 1: Results of quantitative measures.

	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	4.5	1.18	9.8	5.98	6.27	0.001
M3	2.8	3.74	7.8	10.69	6.00	0.001
M4	0.8	1.93	20.3	34.2	10.86	0.001

Figure 5: Results of post-questionnaire.



Getting joint attention responses from Simon helped participants in the with-RJA group have a better idea of whether Simon had learned the concepts or not. This was supported by findings on 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. Also, participants in the without-RJA group requested more confirmations from Simon until they felt confident that Simon learned the concepts (M2). In contrast, in the with-RJA group, participants requested less than six times (i.e., one for each concept), which can be viewed as a baseline. This showed that participants in the with-RJA group had a better understanding of the robot's internal states (i.e., concepts learned or not).

In addition to the quantitative measures, a post-experiment questionnaire shed light on how participants perceived the interaction and the robot. As shown in Figure 5, participants perceived a robot responding to joint attention as more competent and socially interactive. Moreover, results of self-report survey were also consistent with quantitative measures and the questionnaire results. Most participants in the with-RJA group reported that gaze or head orientation were their cues, while participants in the without-RJA group commented that they were frustrated and had tough time telling whether Simon had learned the concepts.

Video analysis of behaviors of participants also confirmed the hypotheses and gave insights to natural HRI. Three interesting observations were found in both groups. First, participants tended to look back and forth between the referred

object and Simon's face (i.e., EJA) to see if Simon understood the concepts. This observation footed the hypothesis that EJA is needed in natural HRI. Second, participants showed RJA (i.e., followed Simon's pointing gesture) when Simon initiated a joint attention (i.e., pointed to the object of request). This observation revealed that it is natural for humans to attend to or respond to joint attention initiated by the other partner. Third, participants tended to give positive or negative responses to the robot. For example, when Simon pointed to the right object, participants nodded, smiled, laughed, or even said positive words, such as "good." Also participants frowned or said negative words such as "no" when Simon pointed to a wrong object. These observations may indicate that humans tend to give responses or feedback to their partners either explicitly (i.e., utterance) or implicitly (i.e., facial expressions).

The Importance of Ensuring Joint Attention in Human-Robot Interaction

We hypothesize that ensuring joint attention (EJA) affects task performance. Moreover, psychological findings (Mundy and Newell 2007; Williams et al. 2005) and our observations in experiment 1 drive us to believe that ensuring joint attention is a natural behavior that humans do. Therefore, we have two hypotheses (H4 and H5) as follows:

- H4: When a robot ensures joint attention it yields better interactive task performance.
- H5: Ensuring joint attention is perceived as a natural behavior in social interaction with a robot.

Experimental design

To test these hypotheses we did a video-based experiment. Participants were given a task to rank a collection of videos where Simon used varying degrees of ensuring joint attention in three different scenarios. The first scenario (presentation) was Simon, as a tour guide robot, giving a presentation to a person. The second scenario (reception) was Simon, as a service robot, received a guest at a reception desk. And the third scenario (directions) was Simon, as a guide robot, directing a person to the restroom in a building.

The presentation scenario was used to verify H4 and H5. The reception scenario was to verify H4, and the directions scenario was to verify H5. In the presentation and reception scenarios, the person in the videos was distracted by some event (i.e., a cell phone call or dropping off a cup of water) during the interaction. Simon responded to the distracting events with four different degrees of EJA in the videos: 1) monitoring and ensuring (m1e1), 2) monitoring but not ensuring (m1e0), 3) no monitoring but ensuring (m0e1), and 4) no monitoring and no ensuring (m0e0).

We use a within-subjects design to measure how people perceive the effectiveness of communication and naturalness of behaviors of a robot in HRI. To minimize order effects on results, we randomly sorted the videos into three different groups (i.e., different orders).

Results

Fifteen participants were recruited for this experiment. All 15 participants (9 males and 6 females) were students from local campus population and were randomly assigned to one of the three groups (five participants in each group). A total of seven participants reported that they had not had experience related to robotics before the experiment.

For both the presentation and the reception scenarios, participants were asked (1) how well the person in the videos can recall or receive the information from the robot and (2) how good the robot was at communicating information. We used the chi-square test for goodness of fit to test participants' first choices to each question. The null hypothesis was that the distribution of people's votes on varying EJA behavior variations is even with respect to those questions. We used the chi-square test to test if the real distribution is significantly different from the even distribution. The significant difference tells us that EJA behaviors would actually affect people's perception of the robot. The results indicated that the full EJA behavior (i.e., m1e1) is the most desirable behavior with respect to the two questions (both significant level at 0.01). The result supported H4 that a robot ensuring joint attention yields better performance in an interactive task. In our scenarios, better performance came from better communications, which is also true in interactive tasks in general.

For both the presentation and the directions scenarios, participants were asked how well Simon engaged the person in the videos. The result showed that the full EJA behavior is the most desirable behavior with respect to engaging the other agent (the chi-square test for goodness of fit, significant level at 0.01). In addition, participants were asked to rank the videos according to how similar the robot's behaviors are to theirs if they were asked to perform the same task. The result revealed that the full EJA behavior is the most similar behavior to theirs (the chi-square test for goodness of fit, significant level at 0.01), suggesting that full EJA is more natural to humans. The results supported H5 that ensuring joint attention is a natural behavior that people do in interaction.

Furthermore, for all three scenarios, participants were asked to rank the videos according to their preference if they were asked to design behaviors for a robot in similar scenarios. For both the presentation and reception scenarios, 14 out of 15 participants agreed that the full EJA behavior is the most desirable one, while 13 participants agreed the same for the directions scenario (the chi-square test for goodness of fit, all significant level at 0.01). This result did not directly support our hypothesis on naturalness of behavior. However, the result that people would like to have EJA behaviors on a robot may imply people perceive EJA behaviors as more affective and natural behaviors.

In the survey of each scenario, participants were asked to comment on the differences they observed and how they liked or disliked the videos. These comments give us insight that it was in fact the ensuring joint attention behaviors that were playing into people's rankings and choices. For the presentation scenario, all participants commented that they noticed a difference where Simon made sure the person

was paying attention before the presentation versus not, such as "waited and made sure that the guest is paying attention before moving on". Twelve participants noted another difference where Simon looked at the user occasionally versus not. Participants often used phrases such as "make eye contact", "engage user", and "recapture attention" to describe the looking back and forth behavior. Similar to the presentation scenario, most participants noticed the two main differences in the reception scenario. Also in the directions scenario, most participants (13) noticed the difference was if the robot turned to the person during interaction. Twelve out of 13 participants commented this behavior in a positive way. 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. In sum, participants' comments on the videos supported data from the survey. Participants preferred behaviors involving ensuring joint attention and believed those are natural behaviors that humans do.

Future Direction

There are still several issues that we have not addressed in current model. First, when and how frequently should a robot need to do ensuring joint attention (EJA) (i.e., monitoring and ensuring) in an interaction? Even though results of the second study has suggested that EJA behavior is natural, we believe that it is natural only when it happens at the time and frequency that meet people's expectation. Second, the model needs to consider dynamics of a crowd to handle interactions with a group of people. We believe that interaction with a person is quite different from interaction 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. More studies are needed to explore dynamics of crowd. 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. Humans have different ways to interact with different people in different situations. To be in a human environment, robots need the ability to adapt themselves.

Conclusion

The contribution of this work is threefold. First, we propose a computational model of joint attention consisting of responding to joint attention, initiating joint attention, and ensuring joint attention. This decomposition is supported by psychological findings and matches the developmental timeline of infancy. Second, in the experiment of exploration of effects of responding to joint attention on human-robot interaction, we found that robots responding to joint attention are more transparent to humans and are more competent and socially interactive. Third, in the experiment of study of the importance of ensuring joint attention in human-robot interaction, we learned that robots ensuring joint attention yield better performance in human-robot interactive tasks and ensuring joint attention behaviors are perceived as natural behaviors by humans.

References

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

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

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

Deák, G. O.; Fasel, I.; and Movellan, J. 2001. The emergence of shared attention: Using robots to test developmental theories. In *In Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, 95–104.

Imai, M.; Ono, T.; and Ishiguro, H. 2003. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transaction on Industrial Electronics* 50(4):636–643.

Kaplan, F., and Hafner, V. V. 2006. The challenges of joint attention. *Interaction Study* 7(2):135–169.

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

Langton, S. R. H.; Watt, R. J.; and Bruce, V. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4(2):50–59.

Mundy, P., and Newell, L. 2007. Attention, Joint Attention, and Social Cognition. *Current Directions Psychological Science* 16(5):269–274.

Nagai, Y.; Y, K. H.; Morita, A.; and Y, M. A. 2003. A constructive model for the development of joint attention. *Connection Science* 15:211–229.

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

Scassellati, B. 1999. 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. 176–195.

Striano, T.; Reid, V. M.; and Hoehl, S. 2006. Neural mechanisms of joint attention in infancy. *The European journal of neuroscience* 23(10):2819–23.

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

Williams, J. H.; Waiter, G. D.; Perra, O.; Perrett, D. I.; and Whiten, A. 2005. An fMRI study of joint attention experience. *NeuroImage* 25(1):133–140.