2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

# Responsive Joint Attention in Human-Robot Interaction

André Pereira, Catharine Oertel, Leonor Fermoselle, Joe Mendelson and Joakim Gustafson

***Abstract*— Joint attention has been shown to be not only crucial for human-human interaction but also human-robot interaction. Joint attention can help to make cooperation more efficient, support disambiguation in instances of uncertainty and make interactions appear more natural and familiar. In this paper, we present an autonomous gaze system that uses multimodal perception capabilities to model responsive joint attention mechanisms. We investigate the effects of our system on people's perception of a robot within a problem-solving task. Results from a user study suggest that responsive joint attention mechanisms evoke higher perceived feelings of social presence on scales that regard the direction of the robot's perception.**

## I. Introduction

Humans are good at observing others and at understanding their mental states, such as what they are seeing, feeling and what they want. An essential social skill required to achieve this is to be able to see and understand the world from someone else's viewpoint. This skill is called Visual Perspective Taking (VPT) [1] and plays a critical role in social cognition and interaction between humans. The same effect has been reported in Human-Robot Interaction (HRI) where people are similarly able to override their egocentric viewpoints and take robots' visual perspectives [2].

If social interaction partners are aware that they are attending something else in common, that phenomenon is usually described as joint attention. Joint attention describes not only the VPT required for this phenomenon but also the tendency for social partners to focus on a common reference and to monitor the others' attention to an outside entity [3]. Joint attention is vital in task-based human-robot collaboration as robots involved in collaborative tasks often need to coordinate their gaze behavior with collaborators and objects in its environment. As an example, a robot in a factory assembly line should use both its verbal and non-verbal joint attention behaviors to reduce uncertainty and better guide human co-workers into which object should be assembled next [4]. Attentional non-verbal behavior generation is not only significant for improving collaboration between people and robots [5] but also for creating the illusion of human-like appearance and behavior.

Most research typically separates joint attention into two different types [6]: Initiating Joint Attention (IJA) and Responding to others' Joint Attention (RJA). The effects of IJA in robots has been extensively explored. To a lesser extent, previous work has demonstrated that robots that respond to others' joint attention are considered as more transparent to humans and perceived as more competent and socially interactive [7]. We intend to extend those findings by studying responsive joint attention not as a means to detect the others' intent but rather with the focus of generating contingent gaze shifts that reduce non-meaningful idle gaze behaviors.

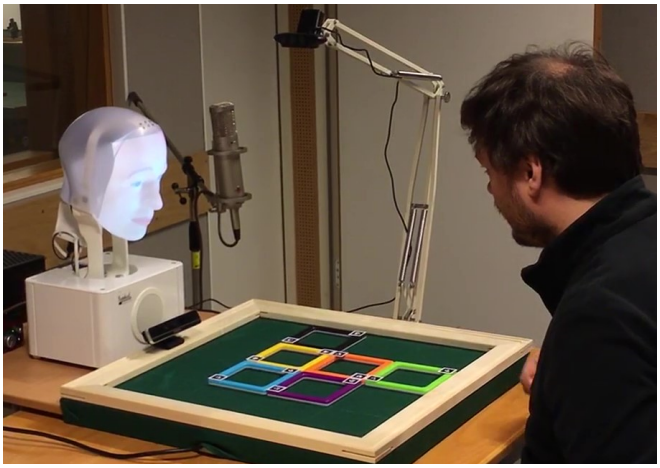

Fig. 1. Side view image of the MagPuzzle case study.

Assuming that people do take robots' visual perspectives, a perception of agency is still inferred from periods where no verbal communication is present. In these moments, robots should display believable non-verbal behaviors that show that they are "alive" and listening which in turn results in an increased sense of social presence [8]. Social presence, often described as the "sense of being together with another" [9], has been shown to increase enjoyment and by consequence improve the intention to use robots over long periods of time [10], [11].

In this paper, we contribute to social presence and joint attention research in human-robot interaction by proposing and evaluating an autonomous real-time system that can generate responsive joint attention by responding to users' gaze direction, speech and actions performed in a task. Our proposed simulation of responsive joint attention is evaluated in a *natural* setting and is not focused on increasing the effectiveness of the user in the interaction but rather on using the perceived multimodal information to minimize idle behaviors and increase the usage of more meaningful joint attention behaviors.

Our system was evaluated within a case study that consists

Andre Pereira, Leonor Fermoselle, Joe Mendelson and Joakim Gustafson are with the Speech, Music and Hearing Lab, EECS at KTH Royal Institute of Technology, Stockholm, Sweden. atap@kth.se, leonor.fermoselle@gmail.com, joemndlsn@gmail.com, jocke@speech.kth.se

Catharine Oertel is with the Computer-Human Interaction Lab for Learning & Instruction at École Polytechnique Fédérale de Lausanne, Switzerland. catharine.oertel@epfl.ch

of a physical spatial reasoning task where a human-like social robot collaborates with humans in finding the solution for a series of puzzles. A user perception study was conducted to test the hypothesis of whether participants perceive our robot as more socially present when our robot simulates responsive joint attention behaviors. Results suggest that the direction of social presence related to the *perception of the robot* is significantly affected by our experimental condition.

## II. Related Work

Body orientation, head, and eye motion are specific non-verbal cues that carry precise meanings in communication. These cues are relevant for establishing and maintaining a connection between two parties as they often reveal a participant's attention. Simulating attention shifts in social robots is fundamental for establishing and driving believable interactions between humans and robots [12].

Joint attention, in particular, is a relevant skill to model in robots as it is widely considered to be a crucial social skill [13], [14], [15]. Interventions focused on developing joint attention will often result in positive collateral changes in untreated skills, such as language and social-cognitive developments. Social robots that simulate joint attention have, for instance, been shown to improve social skills in children on the autism spectrum [16]. Joint attention is also relevant for a human and a robot to coordinate in organizational, assembly and physical tasks. Robots that simulate joint attention mechanisms have been widely shown to help disambiguate spatial references by making object referral appear more natural, pleasant and efficient [17]. Joint attention can also improve the timing and perceived quality of handover events in robot handover tasks [18].

Joint attention can have several phases, beginning with mutual gaze to establish attention, proceeding to referential gaze to draw attention to the object of interest, and cycling back to mutual gaze to ensure that the experience is shared [19]. Different task settings generate different patterns of eye movements and gaze dynamics [20]. As such, collecting eye gaze information in a specific scenario and using it to create gaze models usually produces the best results. These models often make use of stochastic processes and simple statistical distributions [21], [22] or more complex deep learning techniques [23]. Using an exclusive data-driven approach is a costly process that often cannot be extrapolated to other contexts. A different alternative to purely data-driven approaches is to use heuristics to model gaze behavior. Heuristics are usually extracted from gaze focused or psychology literature and are often not bound to a specific scenario. Some gaze modeling approaches use only heuristics to generate gaze shifts to the correct target, with the appropriate length [24], [25], while others use a mixed approach where data collected from the target scenario is used to regulate heuristic processes [21], [22], [26].

## III. Joint Attention System

This section describes a real-time system that responds to users' joint attention by reacting to users' gaze direction, speech and actions performed within a task. Our gaze system considers high-level gaze directives (targets and duration) and does not focus on low-level gaze characteristics (such as eye-head coordination or saccade velocities).

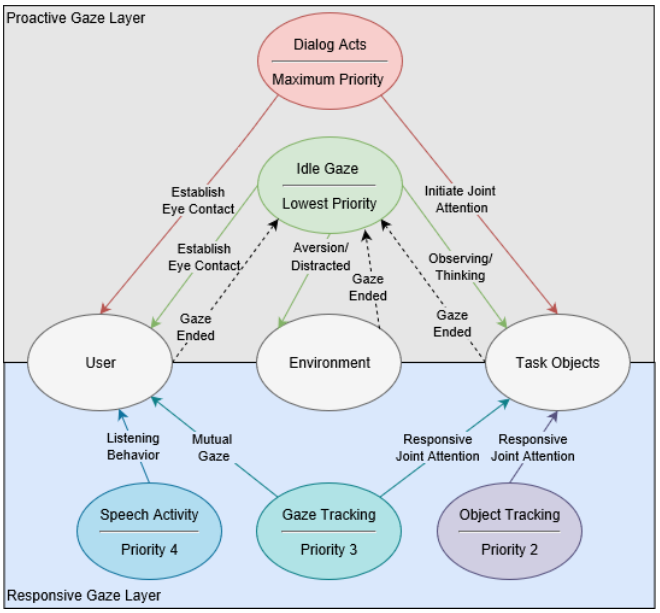


Fig. 2. A diagram illustrating our two concurrent gaze layers (proactive and responsive). Each layer is composed of different modules that can shift the robot's gaze towards the different gaze targets (user, a point regarding a task or object, or non-task points in the environment). Each solid arrow represents a gaze shift to one of the targets in the middle of the diagram. Dashed arrows represent an event that is communicated to our idle gaze module when the current gaze shift issues a time-out.

### A. Layers, Modules, Priorities and Timings

We divide our gaze system into two different layers (see Figure 2). Gaze coordination is known to be a bidirectional process that involves both the production and detection of gaze [21]. Joint attention is also bidirectional in nature and usually divided into two different types (IJA and RJA).

Each layer contains different modules with predefined priority levels. Priorities are defined so that the robot can suppress shifts to irrelevant stimuli and voluntarily hold its attention if currently engaged in a more relevant gaze command. Unless a gaze shift with the same or higher priority is issued, the robot maintains its gaze target until the time duration of the current gaze is completed. In the description for each module, we will explain the reason behind each priority choice.

Regarding timings, in a similar way to [24], [22], [26], [25], [21], we decide for how long a gaze shift will last by using stochastic processes. Every time that a gaze shift is issued, a timer starts with a duration sampled from a continuous uniform distribution within defined intervals. When the timer reaches the target sampled time, a time-out event is triggered. For every module, we started with a time-out range between 0.5 and 1 seconds, an interval recommended by previous literature to convey friendly gaze behavior [24]. However, by following an iterative design ap-

proach, we redefined the values for these parameters at longer intervals (between 1 and 5 seconds). The longer intervals convey a calm, cooperative personality and compensate for the fact that dialog acts and responsive gaze shifts constantly interrupt idle behaviors, hence reducing its average time.

### B. Proactive Gaze Layer

The proactive gaze layer is responsible for gaze commands that are integrated within dialog acts and an idle gaze module that is used to generate idle gaze shifts.

*1) Dialog Acts:* Discrepancies in a robot's gaze and speech may lead to diminished performance which might have a worse effect than not using any gaze model at all [27]. Dialog acts in our system allow for a tight mix of verbal and non-verbal content in a format that is compact for authoring (e.g., see Table I). When a dialog act is selected, one of its multiple content entries is chosen and parsed by a behavior planner, similar to [28], that outputs the right multimodal instructions at the right time to the robot.

Before vocalizing and lip-syncing the text content associated with dialog acts, the robot first shifts its focus of attention towards the intended target defined in the dialog act, and only then executes the associated content. Dialog acts can be associated with the user or objects that belong to the joint attention task. Dialog acts associated with users make the robot attempt to establish and maintain eye contact in real-time until the time-out of the current gaze shift. When people refer to objects belonging to a joint attention task, they often engage in deictic gaze references by looking at those objects ahead of naming or manipulating them [29]. As such, in this type of dialog acts, the robot first attempts to initiate joint attention by looking at the referred object.

This module is defined as having the highest priority which allows it to override any current gaze command when it decides to control the dialog flow.

*2) Idle Gaze:* Idle gaze shifts occur when the timer for any previous gaze shift ends. In these cases, the robot shows that it is alive and takes the proactive stance of performing an idle gaze shift. In these moments, the system either maintains its current gaze target or performs a gaze shift to a different target. Possible targets for idle gaze behavior in joint attention based tasks are typically: looking at users to establish eye contact; at an object related to the task to simulate a thinking or observing behavior; or elsewhere, at a random point in the environment not related to the task or users. A looking elsewhere gaze shift simulates gaze aversion (when shifting from looking at a player) or distraction (when shifting from looking at an object).

Idle gaze shifts are attributed to the lowest priority and can be interrupted by any event from the Dialog Acts module or any module from the Responsive Layer.

### C. Responsive Layer

The sole usage of scripted and randomized behaviors is inadequate for several types of tasks as gaze behavior is often not specified and emergent (related to the dynamics of the environment) [20]. This layer uses data from multimodal perception to minimize the robot's idle time and increase the usage of more meaningful responsive joint attention behavior.

Literature suggests that events not directly related to a task such as peripheral events guide the attention of users. However, if more relevant events are currently engaging the robot's attention, gaze shifts influenced by these events should not occur [20]. Accordingly, gaze shifts triggered by modules in our responsive layer have limited access to gaze control. The robot suppresses shifts to irrelevant stimuli and voluntarily holds its attention if currently engaged in a gaze shift with higher priority. As such, the priority of gaze modules within this layer is recommended to lie in-between dialog generation, and idle gaze modules.

In the remainder of this subsection, we describe and propose three different responsive modules that will be standard in most applications with robots where Joint Attention interaction is a relevant part of the task.

*1) Speech Activity:* It has been shown that people look at the other party in the interaction nearly twice as much (75%) when listening than when speaking (41%) [30], as such, for an added sense of social presence, it is essential that the robot shows that it is an active listener. In our system, when speech activity from a user is detected, the robot looks at the user to simulate listening behavior. This gaze shift happens when the speech recognition stream returns intermediate and final results.

*2) Gaze Tracking:* When a user looks at an object in a task, that information is used to make the robot look at the same objects. This gaze shift is essential to successfully simulate the establishment of responsive joint attention from the robot's perspective. By using gaze tracking that can track task objects, the robot is able to use its perception to understand the user's current focus of attention and match it.

Robots involved in joint attention tasks should also, at appropriate times, switch their gaze to their conversational partners. This coordination is essential for a proper perception of mutual gaze, but it is also a necessary step for guiding the partner's attention and for the establishment of joint attention. Joint attention is better perceived when starting from a mutual gaze phase that establishes shared eye contact before switching attention to a target object. As such, previous establishment of meaningful and timely mutual gaze often contributes to a better perception of joint attention. In addition, favorable feelings for robots are enhanced when eye contact or mutual gaze is used in combination with joint attention [31]. We follow previous findings that recommend agents to engage in mutual gaze when looked at [21].

*3) Object Tracking:* Using real-time data from object tracking to shape the robot's gaze behavior is also essential to respond to the user's attention and establish joint attention from the robot's perspective. When a user moves an object it is highly probable that the user's attention is or will be on the object. In our system, when a user moves, places or removes objects from a task, the robot assumes that the object is the current focus of attention and gazes at that location.

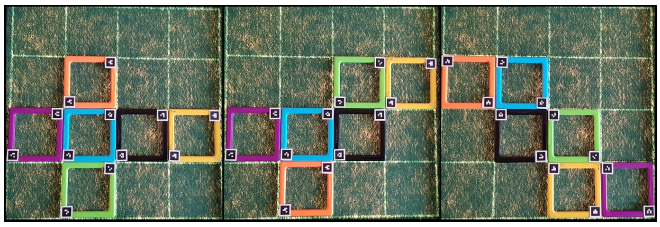

Fig. 3. Example solutions for the X-in-line sub-tasks where X=4 (sub-task 1, image on the left), X=3 (sub-task 2, center) and X=2 (sub-task 3, right).

## IV. MagPuzzle Case study

The model described above was tested within the *MagPuzzle* case study. *MagPuzzle* is a spatial reasoning cooperative task created to study joint attention within a real-time *natural* environment where a life-like social robot interacts with human participants.

### A. Task Rules

In *MagPuzzle*, participants draw the geometry net of a cube by placing and manipulating six different colored puzzle pieces on a two-dimensional board with a 4X4 grid. The task is comprised of three different sub-tasks or puzzles, with increasing levels of difficulty, where the maximum amount of pieces in a line (row or column) allowed for the solution varies (see Figure 3). The social robot introduces the objective of each puzzle and cooperates with participants to arrive at solutions in all three different puzzles. The first puzzle with a maximum of 4-in-line squares is trivial to most participants as it represents the default cube geometry net taught in high-school geometry. However, most participants struggle with the second and third puzzles, which gives the robot opportunities for offering assistance while participants are pondering the possible moves, thereby increasing the chance of the robot to socially engage and simulate joint attention behaviors. The robot guides users toward a solution by using non-deterministic speech combined with joint attention gaze.

### B. Natural Environment Setup

Our definition of a *natural* environment comprises three different elements:

First, we are using a robotic head that is capable of inducing and responding to joint attention directives through eye gaze, head pose direction, facial expressions, and verbal cues. The robot used in our setup is a Back-projected head [32] placed opposite to participants next to a table that contains the puzzle board (see Figure 1). We use a back-projected robotic head as they allow for very fine-grained gaze-driven deictic gestures and have been shown to be capable platforms for allowing gaze reading by humans [33]. Our robot is capable of digitally animating quick eye gaze changes and uses two degrees of freedom servos to simulate head movements. The robot height was adjusted to make the interaction comfortable and accommodating to most heights.

Second, we are using non-intrusive technology to sense speech, gaze direction, and actions performed by users. In front of the robot, we are using an RGB-D camera that is used for user and gaze tracking. For object tracking, we are using a dedicated RGB camera on top of the board that detects fiducial markers placed on the puzzle pieces. For speech activity detection and speech recognition, we are using a high-quality amplified mono-directional microphone placed on the side of the robot. One of the key aspects we focused on while designing the hardware setup was to keep it simple and non-intrusive. In contrast to many studies in this area, users were not requested to wear glasses, participate in any calibration step or wear dedicated microphones. These decisions may hinder the performance of our perception modules but more importantly, do not influence the flow and naturalness of the experience.

Finally, our task was purposefully designed so that the robot's guidance is more successful if joint attention is established with the robot, but also that the robot is not essential for the completion of the task. Participants can choose to cooperate with the robot but can also choose to ignore it and still be successful. This setting allows us to better tackle our research goals as the simulation of responsive joint attention is not necessarily going to improve task efficiency but rather the subjective perception of our robot.

### C. Multimodal Perception

The *MagPuzzle* system[1] can sense various multimodal cues from the participants. Our system can recognize participants' speech, their head orientation and can detect pieces that are placed or removed from the board. The multimodal perception information is used to generate the responsive gaze behavior described in the previous section. In this subsection, we briefly describe our three different perception modules.

For *Speech Activity* detection, we are using Microsoft's commercial cloud speech recognition service, a service that supports real-time streaming. When the speech recognition system detects a start of speech event we use that event as a reliable detection of speech activity.

For *Gaze Tracking*, we are using GazeSense[2], a non-intrusive gaze tracker, used to estimate participants' visual focus of attention. This system supports two modes of detecting gaze direction. A simple model that uses head orientation data as input, and a more precise mode that combines that information with eye pupil data. In our case, a single camera was not adequate to precisely track the eye's pupil when the gaze is directed towards a horizontal surface. As such, we decided to only use head orientation for gaze tracking. This decision is backed by research showing that head orientation contribution in predicting overall gaze direction is estimated to be around 70% [34]. Several other systems that focus on real-time responsive behavior generation in socially interactive agents [35], [36], [25], [37], [21], have reported good levels of accuracy by relying solely on head

[1]Code available on `https://github.com/andre-pereira/MagPuzzle`

[2]https://eyeware.tech/gazesense/

TABLE I

SHORT ILLUSTRATION OF THE DIALOG ACTS THAT WERE MOST COMMONLY USED IN *MagPuzzle*.

| Dialog Act | Content | Target |
|---|---|---|
| Remind Objective | Remember {user_name}, we will need to build 4 walls a roof and a floor. | Player |
| Well Played | I could not have thought <gesture(express_happy)> of a better move myself. | Player |
| Hint | I think putting a piece here might help. | Square |
| Wrong Piece | I would think of a different place for this piece. | Piece |

pose to determine humans' focus of attention. Nevertheless, given the limitations in the performance of head orientation detection, we could not distinguish individual squares in our board but instead, four different quadrants each representing four squares (top-left, top-right, bottom-left, bottom-right). In total, our gaze tracking system can detect if users' are gazing at six different targets (the robot, any of the four quadrants or somewhere else).

For *Object Tracking* (tracking the *MagPuzzle* pieces), we integrated ARToolkit[3] in our system, an open-source augmented reality toolkit that can, in real-time, robustly recognize and track multi-marker fiducials placed on the puzzle pieces used in our scenario.

### D. Dialog Management

For dialog management and dialog act selection we use a human-in-the-loop approach where a human operator does not directly control the gaze behavior of the robot but is responsible for selecting the appropriate dialog act at the right moment. A total of 36 dialog acts were created for *MagPuzzle*. To prevent the repetition of the same dialog act, each had multiple content implementations, and a total of 234 lines of authored content were created. Dialog acts are used to: describe and manage the rules of the task (21); provide simple responses to users' questions (3); offer hints on what to play next (3); provide feedback on the last move (2); provide feedback on the overall state of the board (2); provide feedback on a specific square on the board (2); and to motivate or compliment the user (3).

### E. Hints and Feedback

We developed a helper search algorithm (brute force, depth-first) that helps the human-in-the-loop to know: the correct squares for providing hints to the user; if the board is in a proper or incorrect board state; or if any rule was broken. This algorithm avoids mistakes and helps in automating the choice of appropriate hint dialogue acts. Verbal hints provided by the robot, are accompanied by a gaze shift to the right square. Thus, hints are not explicit and indecipherable if the user does not follow the robot's attempt to initiate joint attention.

## V. EVALUATION

This section presents details from a user study performed to evaluate our system and hypothesis.

[3]https://github.com/artoolkit/artoolkit5

### A. Study Objectives

Our user study had the main objective of testing the hypothesis of whether the participants in the full condition perceive the robot as more socially present. In addition, we will perform a manipulation check that confirms the validity of our conditions and we look at the success rate of the hints suggested by the robot.

### B. Manipulation

We created two different conditions to evaluate the effects of using multimodal perception to generate responsive joint attention behaviors:

- **full -** our experimental condition uses the full joint attention system described in section III.
- **proactive -** in this condition, the gaze system is only composed of the proactive layer (proactive) and does not use the responsive layer modules to generate gaze shifts (see Figure 2).

We designed and developed the proactive layer and condition with the best behavior we could achieve without using the responsive layer with the goal of making a fair comparison between both conditions.

### C. Participants

Twenty-two participants (11 full, 11 proactive) took part in a 30-45 minute between-subjects experiment in a lab experiment at KTH, Royal Institute of Technology in Sweden, Stockholm. Participants were aged 18-34 years old and were rewarded with a cinema ticket.

### D. Procedure

At the beginning of the experiment, participants were guided to a room that contained our setup and asked to fill a consent form. Next, participants were randomly assigned to one of our conditions and debriefed by the experimenter. At the beginning of each interaction, the robot explains the MagPuzzle task in detail and participants complete the set of three puzzles in a time span that lasted from 5-10 minutes depending on the participant. Finally, participants fill in a questionnaire, and the experiment is concluded.

### E. Measures

*1) Social Presence Questionnaire:* The relationship between the perception of social presence and joint attention is still quite unexplored in HRI. However, we strongly advocate the use of social presence measures to evaluate the believability of joint attention systems in social robots. The bidirectional quality of this measure makes it a sound choice

TABLE II

SOCIAL PRESENCE ITEMS USED FOR EACH DIRECTION OF SOCIAL PRESENCE AND FOR EACH OF ITS SIX DIMENSIONS [38].

| | **Perception of self** | **Perception of the robot** |
|---|---|---|
| Co-Presence | I noticed the robot.<br>The robot's presence was obvious to me.<br>The robot caught my attention. | The robot noticed me.<br>My presence was obvious to the robot.<br>I caught the robot's attention. |
| Attentional Allocation | I was easily distracted from the robot when other things were going on.<br>I remained focused on the robot throughout our interaction.<br>The robot did not receive my full attention. | The robot was easily distracted from me when other things were going on.<br>The robot remained focused on me throughout our interaction.<br>I did not receive the robot's full attention. |
| Message Understanding | The robot's thoughts were clear to me.<br>It was easy to understand the robot.<br>Understanding the robot was difficult. | My thoughts were clear to the robot.<br>The robot found it easy to understand me.<br>The robot had difficulty understanding me. |
| Behavioral Interdependence | My behavior was often in direct response to the robot's behavior.<br>I reciprocated the robot's actions.<br>My behavior was closely tied to the robot's behavior. | The behavior of the robot was often in direct response to my behavior.<br>The robot reciprocated my actions.<br>The robot's behavior was closely tied to my behavior. |
| Emotional Understanding | I could tell how the robot felt.<br>The robot's emotions were not clear to me.<br>I could describe the robot's feelings accurately. | The robot could tell how I felt.<br>My emotions were not clear to the robot.<br>The robot could describe my feelings accurately. |
| Emotional Interdependence | I was sometimes influenced by the robot's moods.<br>The robot's feelings influenced the mood of our interaction.<br>The robot's attitudes influenced how I felt. | The robot was sometimes influenced by my moods.<br>My feelings influenced the mood of our interaction.<br>My attitudes influenced how the robot felt. |

for evaluating the effects of each type of joint attention individually (IJA and RJA). Social presence measures the degree to which an individual feels interconnected with another entity [39]. To achieve a feeling of interconnection, participants should believe that both the robot's behavioral and psychological engagement is connected to theirs and that their behavioral and psychological engagement is connected to the robot. For measuring social presence, we are using a social presence questionnaire that divides social presence into six different dimensions [38]. Similarly to [40], we divide each dimension in two distinct directions that measure the social presence of the participant in connection with the robot (*perception of self*) and three questions that measure the perception of the robot in connection with the participant (*perception of the robot*). We used all 36 items from the original questionnaire in a 5 point Likert-scale format (see Table II).

*2) Objective Data:* In addition to the questionnaires, we video recorded the interactions and logged all of the information collected by the joint attention system, including all gaze shifts from the participants, and the robot, the moves performed in the board and the selected dialog acts. This information will be used to: present gaze distributions that reflect the robot's behavior and its synchrony with participants; perform a manipulation check; and to present results on hint success rate.

### F. Results

*1) Gaze Distributions:* We analyzed all the gaze shifts of the robot to better illustrate our gaze system. Table III shows the percentage of time that the robot spent on average in each module and gaze target divided by condition. Similarly to [37], we look at the synchrony between participants' gaze behavior and the robot. For this particular analysis (see Table IV), we only considered N=5 participants from the proactive condition and N=8 participants from the full condition, a consequence of not logging all of the participants' gaze data in the initial sessions. This later analysis is not essential to our research question and is only used for our manipulation check, hence the decision of not manually annotating the missing data.

TABLE III

AVERAGE TIME, PER CONDITION, FOR EACH TYPE OF GAZE.

| | | Full | Proactive |
|---|---|---|---|
| Dialog Acts | Player | 33,61% | 33,80% |
| | Board | 13,54% | 16,10% |
| Idle Gaze | Player | 17,95% | 37,26% |
| | Board | 2,61% | 11,74% |
| | Elsewhere | 0,34% | 1,10% |
| Speech Activity | Player | 3,16% | N/A |
| Gaze Tracking | Player | 4,68% | N/A |
| | Board | 7,01% | N/A |
| Object Tracking | Board | 17,08% | N/A |
| TOTAL | Player | 59,41% | 71,06% |
| | Board | 40,25% | 27,84% |

TABLE IV

AVERAGE TIME, PER CONDITION, FOR BEHAVIORAL SYNCHRONY.

| Condition | Mutual Gaze | Joint Attention | Mismatch |
|---|---|---|---|
| full | 41,69% | 22,51% | 35,8% |
| proactive | 44,99% | 7,58% | 47,46% |

*2) Manipulation Check:* As a manipulation check, we look into moments of where the the robot establishes joint attention with the user (from its own perspective). The average time of responsive joint attention significantly increased in the full condition (M=22.51, SD=15.44) compared to proactive (M=7.58, SD=6.48); $t(11)=-2.415$, $p<.05$. As anticipated, given our manipulation, these results show the difference in the robot's gaze behavior between both conditions and reflect the increased and timely joint attention to the board (pieces and quadrants).

*3) Hints Success Rate:* The hints provided by the robot during the task depend on the participant's ability to adopt the visual focus of attention of the robot and perceive the game from the robot's perspective. We observed that participants could indeed successfully take the robots visual

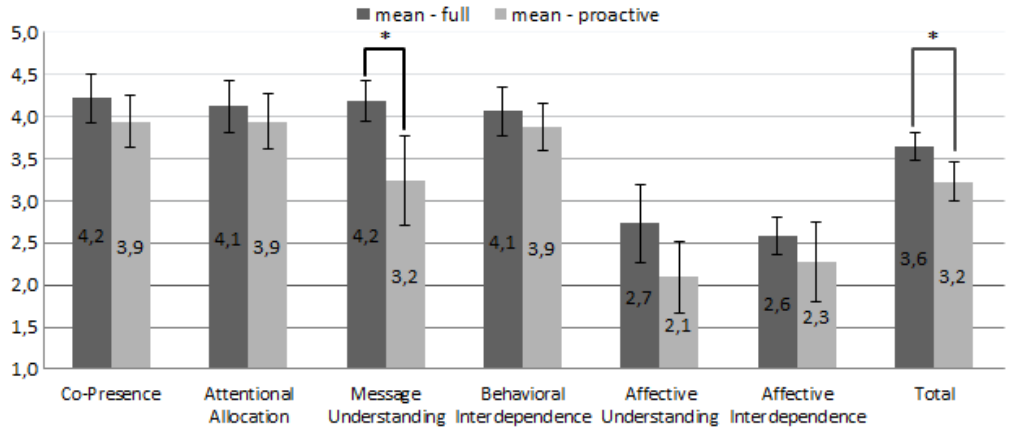


Fig. 4. *Perception of the robot* mean values of the social presence measures.

perspective, because the hints and feedback provided by the robot were most often correctly assessed. The hints provided by the robot regarding the next move for the puzzle are given by head pose direction and eye gaze combined with an indirect verbal reference (e.g. "I think putting a piece *here* might help"). Although this type of cue was difficult to perceive, given the resolution of the board with pieces placed close together, the success rate of these hints was still 65% at the first prompt. This rate shows that participants had a good understanding of the robot's viewpoint of the board. Nonetheless, the success rate of combined verbal and non-verbal feedback cues was clearer and reached a 93% success rate (e.g., I would arrange this yellow piece differently).

*4) Social Presence Questionnaire:* Cronbach's $\alpha$ tests confirmed the validity of the social presence questionnaire and revealed good to excellent reliability scores in all six dimensions for both social presence directions. Figure 4 presents the mean results from the items in the *perception of the robot* direction (no significant differences were found in the *perception of self* direction). A Mann-Whitney test indicated that the total of *perception of the robot* was significantly higher in the full condition (Mdn = 3.6) than in the proactive condition (Mdn = 3.1) $U$ = 27.5, p = .015, r = .46. Message understanding ($U$ = 29.0, p = .018, r = .45) was the dimension most responsible for this effect. The most revealing item from this dimension clearly shows the effect of our responsive joint attention system, "My thoughts were clear to the robot" ($U$ = 12.0, p = .001, r = .70).

## VI. Conclusions

We present a gaze system that generates responsive joint attention in human-robot interaction. In addition to initiating joint attention processes, our system autonomously responds to participants' multimodal cues that reveal their attention. The perception of these cues enables a reduction in the time that the robot is engaged in random idle behaviors and allows it to exhibit more meaningful and responsive gaze shifts.

The experiment presented in this paper shows the importance of enabling robots with joint attention mechanisms that are not necessarily relevant for the completion of cooperative tasks and exemplify the benefits of dividing the social presence measure into two different directions for measuring bidirectional aspects of joint attention in social robots. In a between subjects experiment, participants perceived the robot in the full condition as more socially present in the measures related to the *perception of the robot* social presence direction. The questionnaire items in this direction require users to employ their theory of mind [41] to mind-read [42] the robot and assess its perception capabilities. As such, given that we were manipulating the social robot's perceptive and responsive capabilities we expected to find stronger effects of our manipulation between conditions regarding the second-order construct. To a lesser extent, we also expected to improve the items in the *perception of self* direction, but no significant differences were found. We hypothesize that these measures are more dependent on other factors such as providing more tangible help or by manipulating dialog acts. However, manipulating these factors was outside the scope of our experiment and proving this hypothesis would require additional studies.

Our modular proposal for joint attention simulation includes a responsive layer that uses non-intrusive technology, and that can adapt to most HRI scenarios where joint attention is relevant. Autonomous generation of responsive joint attention behaviors can be valuable for both human-in-the-loop studies and completely autonomous systems. Typically, any scenario where people and robots naturally interact on a task over a shared visual space would benefit from believable responsive gaze behavior. Social robots that interact in long-term or entertainment applications are perfect applications for our proposed joint attention mechanisms.

## Acknowledgment

The work has been funded by the BabyRobot project, supported by the EU Horizon 2020 Programme, grant 87831.

## References

[1] J. H. Flavell, "The development of knowledge about visual perception." in *Nebraska symposium on motivation*. University of Nebraska Press, 1977.

[2] X. Zhao, C. Cusimano, and B. F. Malle, "Do people spontaneously take a robot's visual perspective?" in *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 2016, pp. 335–342.

[3] M. Tomasello *et al.*, "Joint attention as social cognition," *Joint attention: Its origins and role in development*, vol. 103130, 1995.

[4] C. Lenz, S. Nair, M. Rickert, A. Knoll, W. Rosel, J. Gast, A. Bannat, and F. Wallhoff, "Joint-action for humans and industrial robots for assembly tasks," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 2008, pp. 130–135.

[5] H. Admoni, T. Weng, B. Hayes, and B. Scassellati, "Robot nonverbal behavior improves task performance in difficult collaborations," in *The 11th ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 51–58.

[6] J. N. Meindl and H. I. Cannella-Malone, "Initiating and responding to joint attention bids in children with autism: A review of the literature," *Research in developmental disabilities*, vol. 32, no. 5, pp. 1441–1454, 2011.

[7] C.-M. Huang and A. L. Thomaz, "Joint attention in human-robot interaction." in *AAAI Fall Symposium: Dialog with Robots*, 2010.

[8] A. Pereira, R. Prada, and A. Paiva, "Socially present board game opponents," in *Advances in Computer Entertainment*. Springer, 2012, pp. 101–116.

[9] F. Biocca, J. Burgoon, C. Harms, and M. Stoner, "Criteria and scope conditions for a theory and measure of social presence," *Presence: Teleoperators and virtual environments*, 2001.

[10] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Influence of social presence on acceptance of an assistive social robot and screen agent by elderly users," *Advanced Robotics*, vol. 23, no. 14, pp. 1909–1923, 2009.

[11] I. Leite, C. Martinho, A. Pereira, and A. Paiva, "As time goes by: Long-term evaluation of social presence in robotic companions," in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE, 2009, pp. 669–674.

[12] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 2010, p. 5.

[13] R. L. Koegel and W. D. Frea, "Treatment of social behavior in autism through the modification of pivotal social skills," *Journal of Applied Behavior Analysis*, vol. 26, no. 3, pp. 369–377, 1993.

[14] P. Mundy and M. Crowson, "Joint attention and early social communication: Implications for research on intervention with autism," *Journal of Autism and Developmental disorders*, vol. 27, no. 6, pp. 653–676, 1997.

[15] E. A. Jones and E. G. Carr, "Joint attention in children with autism: Theory and intervention," *Focus on autism and other developmental disabilities*, vol. 19, no. 1, pp. 13–26, 2004.

[16] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with asd using a long-term, in-home social robot," *Science Robotics*, vol. 3, no. 21, p. 7544, 2018.

[17] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. André, "Exploring a model of gaze for grounding in multimodal hri," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 247–254.

[18] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft, "Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing," in *Proceedings of the 2014 ACM/IEEE International conference on Human-robot interaction*. ACM, 2014, pp. 334–341.

[19] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[20] S. C. Khullar and N. I. Badler, "Where to look? automating attending behaviors of virtual human characters," *Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1-2, pp. 9–23, 2001.

[21] S. Andrist, M. Gleicher, and B. Mutlu, "Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters," in *CHI*, 2017.

[22] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater, "An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience," in *Computer Graphics Forum*, vol. 23, no. 1. Wiley Online Library, 2004, pp. 1–11.

[23] K. Stefanov, "Recognition and generation of communicative signals: Modeling of hand gestures, speech activity and eye-gaze in human-machine interaction," Ph.D. dissertation, KTH Royal Institute of Technology. Ph.D. Thesis, 2018.

[24] A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita, "Messages embedded in gaze of interface agents—impression management with agent's gaze," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2002, pp. 41–48.

[25] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 67–74.

[26] A. Pereira, R. Prada, and A. Paiva, "Improving social presence in human-agent interaction," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 1449–1458.

[27] M. Staudte and M. W. Crocker, "Visual attention in spoken human-robot interaction," in *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*. IEEE, 2009, pp. 77–84.

[28] T. Ribeiro, A. Pereira, E. Di Tullio, P. Alves-Oliveira, and A. Paiva, "From thalamus to skene: High-level behaviour planning and managing for mixed-reality characters," in *IVA 2014 Workshop on Architectures and Standards for IVAs*, 2014.

[29] Z. M. Griffin and K. Bock, "What the eyes say about speaking," *Psychological science*, vol. 11, no. 4, pp. 274–279, 2000.

[30] R. Vertegaal, R. Slagter, G. Van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001, pp. 301–308.

[31] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe, "Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking," in *Proceedings of the 9th International conference on Multimodal interfaces*. ACM, 2007, pp. 140–145.

[32] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*. Springer, 2012, pp. 114–130.

[33] F. Delaunay, J. de Greeff, and T. Belpaeme, "A study of a retro-projected robotic face and its effectiveness for gaze reading by humans," in *Proceedings of the 5th ACM/IEEE International conference on Human-robot interaction*. IEEE Press, 2010, pp. 39–44.

[34] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*. IEEE, 2002, pp. 273–280.

[35] C. Peters, S. Asteriadis, and K. Karpouzis, "Investigating shared attention with a virtual agent using a gaze-based interface," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 119–130, 2010.

[36] M. M. Hoque and K. Deb, "Robotic system for making eye contact pro-actively with humans," in *Electrical & Computer Engineering (ICECE), 2012 7th International Conference on*. IEEE, 2012, pp. 125–128.

[37] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to with-me-ness in human-robot interaction," in *The 11th ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 157–164.

[38] C. Harms and F. Biocca, "Internal consistency and reliability of the networked minds measure of social presence," 2004.

[39] F. Biocca, C. Harms, and J. Gregg, "The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity," in *4th annual International workshop on presence, Philadelphia, PA*, 2001, pp. 1–9.

[40] F. Biocca and C. Harms, "Guide to the networked minds social presence inventory v. 1.2: Measures of co-presence, social presence, subjective symmetry, and intersubjective symmetry," *Michigan State University, East Lansing*, 2003.

[41] C. Frith and U. Frith, "Theory of mind," *Current Biology*, vol. 15, no. 17, pp. R644–R645, 2005.

[42] S. Nichols and S. P. Stich, *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press, 2003.