A Novel Architectural Method for Producing Dynamic Gaze Behavior in Human-Robot Interactions

Gordon Briggs
Code 5512
U.S. Naval Research Laboratory
Washington, DC USA
gordon.briggs@nrl.navy.mil

Meia Chita-Tegmark
Human-Robot Interaction Laboratory
Tufts University
Medford, MA USA
Mihaela.Chita_Tegmark@tufts.edu

Evan Krause
Human-Robot Interaction Laboratory
Tufts University
Medford, MA USA
evan.krause@tufts.edu

Will Bridewell
Code 5512
U.S. Naval Research Laboratory
Washington, DC USA
ORCID: 0000-0003-3676-9279

Paul Bello
Code 5512
U.S. Naval Research Laboratory
Washington, DC USA
paul.bello@nrl.navy.mil

Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA USA
matthias.scheutz@tufts.edu

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 978-1-6654-0731-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/HRI53351.2022.9889499

***Abstract*—We present a novel integration between a computational framework for modeling attention-driven perception and cognition (ARCADIA) with a cognitive robotic architecture (DIARC), demonstrating how this integration can be used to drive the gaze behavior of a robotic platform. Although some previous approaches to controlling gaze behavior in robots during human-robot interactions have relied either on models of human visual attention or human cognition, ARCADIA provides a novel framework with an attentional mechanism that bridges both lower-level visual and higher-level cognitive processes. We demonstrate how this approach can produce more natural and human-like robot gaze behavior. In particular, we focus on how our approach can control gaze during an interactive object learning task. We present results from a pilot crowdsourced evaluation that investigates whether the gaze behavior produced during this task increases confidence that the robot has correctly learned each object.**

I. Introduction

In our day-to-day lives, we constantly shift our attentional focus to different locations in the outside world and to our own mental processes. Interpreting and reacting to signals of others' attentional focus is an essential part of social interaction [20]. Gaze, being one of the principle signals of attentional focus, has been a major topic of research for enabling more natural and effective human-robot interactions [2], [52]. Gaze is also a window into what tasks people are performing [50], what objects they are thinking about [44], and what mental strategies they may be using to achieve their tasks [23]. Therefore, developing computational methods for robots to produce and reciprocate gaze behavior is important for multiple reasons. Natural and human-like gaze behavior may facilitate human-robot social interactions. Additionally, enabling robots to engage in information processing during tasks in more human-like, attention-bound ways, and using gaze to reflect this processing, could be one method of producing more *explainable* robotic behavior [19].

Over the years, researchers have developed a wide range of computational methods for generating gaze behaviors in artificial agents (see [2] for a survey). These range from data-driven and heuristic approaches that produce gaze behavior tailored for a specific task (either trained or engineered, respectively), to biologically inspired approaches that use a more general, human-inspired model of visual attention or cognition as a basis for producing gaze behavior. These "biologically-inspired" approaches range from neuro-biologically inspired models of human visual attention [24] to higher-level cognitive architectures such as Soar [27] or ACT-R [6].

While biologically inspired approaches to generating gaze behavior may generalize to a wider set of tasks and contexts, they still have some key limitations. Neuro-biological approaches, such as those in the Itti and Koch tradition [24], provide high-fidelity models of some aspects of human visual attention, but they fail to account for how attention can be captured and directed toward other non-visual activities and cognitive processes. In contrast, while cognitive architectures such as Soar and ACT-R provide a more general account of human cognitive processes, none of these architectures have a dedicated core mechanism for modeling the role of attention in these processes.

In this paper, we present a novel integration of a computational framework for modeling attention-driven processes, ARCADIA [14], with a cognitive robotic architecture, DIARC [43]. In contrast to prior biologically inspired approaches, ARCADIA provides a novel framework with an attentional mechanism that bridges both lower-level visual and higher-level cognitive processes. We demonstrate this integration in different tabletop scenarios. In particular, we focus on the

application of the integrated system to control gaze behavior during an interactive learning task. In this task, a human teaches a robot the names of novel objects, then asks the robot to point to a particular object. To evaluate whether the gaze behavior of the system improves human ratings of confidence that the robot has learned correctly, we conducted a pilot crowdsourced evaluation. We report on initial results from this pilot and discuss the strengths and limitations of the current evaluation. Finally, we discuss the potential of this integration and directions for expanded capabilities.

II. Architecture Integration

We present a new approach to generating gaze behavior in robotic agents by integrating two existing architectures with complementary abilities: (1) the ARCADIA architecture, which provides a framework for modeling attentional processes; and (2) the DIARC architecture, which provides a flexible framework to enable a wide-range of cognitively-inspired behaviors on robotic platforms.

A. ARCADIA

ARCADIA provides a computational framework in which *attention* is a central organizing mechanism that unifies perceptual, cognitive, and action-oriented processes [14]. An instance of ARCADIA consists of a few key elements: a set of information processing *components*, a central information buffer called *accessible content*, an *attentional strategy*, and a *focus of attention*. Additionally, an instance of ARCADIA also consists of an *environment* and *sensors* that take information from the environment and push it to accessible content.

ARCADIA operates in discrete computational cycles. During each cycle, each component reads information from accessible content and produces new output that will constitute the contents of accessible content during the next cycle. Furthermore, on each cycle, the attentional strategy selects an information element from accessible content to be the new focus of attention during the next cycle. The current focus of attention is available to each component, which may be responsive to the focus or operate independently. In this way, ARCADIA has a natural mechanism to model processes that require attentional focus and ones that do not. In addition, there are no representational constraints within each ARCADIA component, though information sent to accessible content is packaged in a common format. Discrete packages of information in accessible content are referred to as *interlingua elements*.

Because of this representational flexibility and ability to model both attention-bound and attention-independent processes, ARCADIA has been used to successfully model a variety of phenomena from human psychology. For example, ARCADIA can model visual perception tasks such as multiple-object tracking [29] and numerical perception [16]. ARCADIA also can model phenomena that bridges both perception and more complex, cognitive tasks, such as identifying causal relationships between events [10]. However, these previous models all relied on input from video files or still images. To enable ARCADIA to both receive sensory input from physical sensors and control physical affectors, we aimed to integrate the system with a pre-existing cognitive robotic architecture.

B. DIARC

The "Distributed Integrated Affect Reflection Cognition" (DIARC) architecture is a cognitive architecture for embodied agents [43]. DIARC is implemented in ADE, a middleware infrastructure [38], [40]. The ADE implementation of DIARC provides an open component-based and modular design via defined APIs that can be easily extended to integrate new components or connect to other architectures [21]. In ADE, a series of DIARC components engage in independent processing and asynchronous communication. DIARC components can be also distributed across multiple computers. Communication between components is facilitated by the `ADE-Registry`, which tracks the status and network location of each component.

The DIARC architecture has been used to enable multimodal human-robot interactions through speech, text, and GUIs [32], [37]. In order to enable better coordination with humans, DIARC possesses deeply integrated natural language capabilities [17], [49], including the ability to learn novel objects and action through dialogue [41]. Additionally, DIARC also includes the ability to reason about theory of mind and shared mental models [18], [22], [39]. However, one feature to improve the quality of human-robot interactions that DIARC has lacked previously is a systematic mechanism to produce gaze behavior.

C. Core ARCADIA/DIARC Integration

Figure 1 depicts the key components from both systems used in the integration. We created a new DIARC component, `ArcadiaComponent`, that serves as a wrapper for an instance of ARCADIA. This component is responsible for collecting information from other DIARC components and communicating it to the ARCADIA instance. While DIARC itself is an asychronous architecture, each DIARC component operates on fixed computational cycles. We tied each ARCADIA cycle to the update cycle of the wrapper component. During each `ArcadiaComponent` cycle, information from DIARC is packaged and placed in ARCADIA's environment buffer. Additionally, the `ArcadiaComponent` obtains the current focus of attention selected by ARCADIA's attention strategy and generates commands to move the robot's head orientation conditioned on the current focus of attention.

D. ARCADIA/DIARC Components

While the `ArcadiaComponent` is responsible for communicating with the instance of the ARCADIA, individual DIARC and ARCADIA components are responsible for producing, consuming, and repackaging specific informational elements. For instance, when a robot is situated in front of multiple objects, DIARC's vision system may detect these objects, creating corresponding DIARC `MemoryObject` representations. Dialogue history that may contain linguistic

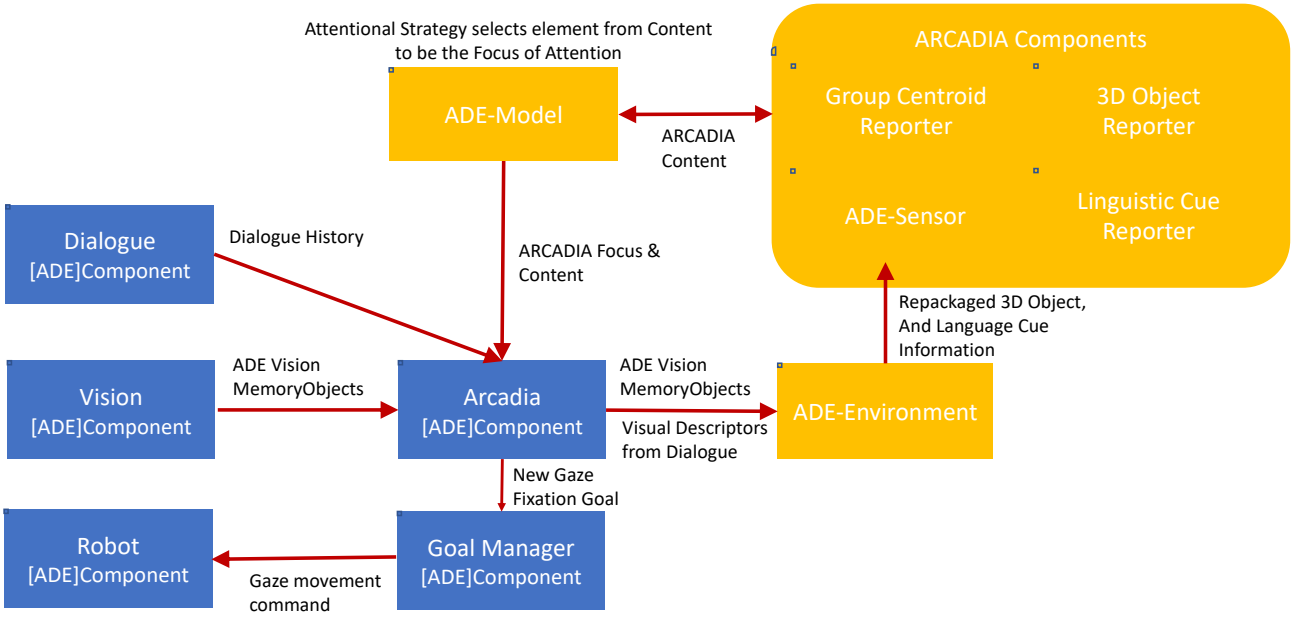

Fig. 1. Diagram depicting the system components involved in the ARCADIA/DIARC integration. Blue represents architectural features and components from the DIARC architecture, while yellow represents architectural features and components from ARCADIA.

descriptors matching visual objects is also collected from DIARC's dialogue component. These DIARC objects are sent to ARCADIA's environment buffer. An ARCADIA sensor takes DIARC objects from the environment buffer and repackages them as ARCADIA elements. In the case of a DIARC `MemoryObject`, an ARCADIA 3D object reporter component wraps `ade-memory-object`, pushing them to ARCADIA's accessible content buffer. The ARCADIA linguistic cue reporter takes the dialogue information from DIARC and produces an ARCADIA `linguistic-cue` element, if the dialogue information contained any visual description predicates.

ARCADIA's attentional strategy then selects an element from accessible content to be the focus of attention. The DIARC `ArcadiaComponent` then queries the current focus of attention from the ARCADIA instance. The `ArcadiaComponent` contains conditional code that determines when a gaze orientation command should be issued and the new gaze direction. Gaze shift timing can be influenced by multiple factors such as a typical gaze shift delay parameter τ_{idle}, which dictates baseline frequency of gaze shifts, or special cases that may preempt the typical gaze shift frequency. For example, if the robot hears a new utterance (and an updated dialogue history is retrieved by the `ArcadiaComponent`), a gaze shift is initiated faster than the baseline frequency. Additionally, the `ArcadiaComponent` contains code that conditionally responds based on the type of ARCADIA interlingua element retrieved as the focus of attention. Currently, if the current focus is an `ade-memory-object` element, then the original DIARC `MemoryObject` is retrieved and a goal to orient the head pose to gaze at this object is submitted to the DIARC `GoalManagerComponent`. While the system is presently only responsive to `ade-memory-object` elements, in principle a wide-range of different gaze shifting behaviors could be developed that differ based on the retrieved ARCADIA focus of attention.

III. Attentional Strategies

In the previous section, we introduced the ARCADIA and DIARC architectures and described how these systems were integrated. However, we have not yet addressed how this integration enables the generation of useful robotic gaze behavior. As mentioned previously, the `ArcadiaComponent` generates a command to move the robot's head orientation based on the focus of attention from ARCADIA. However, what determines the focus of attention is ARCADIA's current *attentional strategy*.

The concept of an attentional strategy is an intuitive one. Imagine you are an office worker tasked with three responsibilities:

- (A) Process paperwork
- (B) Respond to emails
- (C) Perform online training

You will likely come up with a strategy on how to prioritize these tasks. For instance, your strategy might be: $B \succ A \succ C$. In this case, you immediately respond to emails as they arrive. In the absence of new emails, though, you do not simply sit idle. Rather, you check to see if there are any outstanding pieces of paperwork that need to be completed. Likewise, if no new pieces of paperwork are available, you then see if you have any online training to do.

But, what if you are suddenly faced with an urgent deadline to complete mandatory training courses? Then, you might

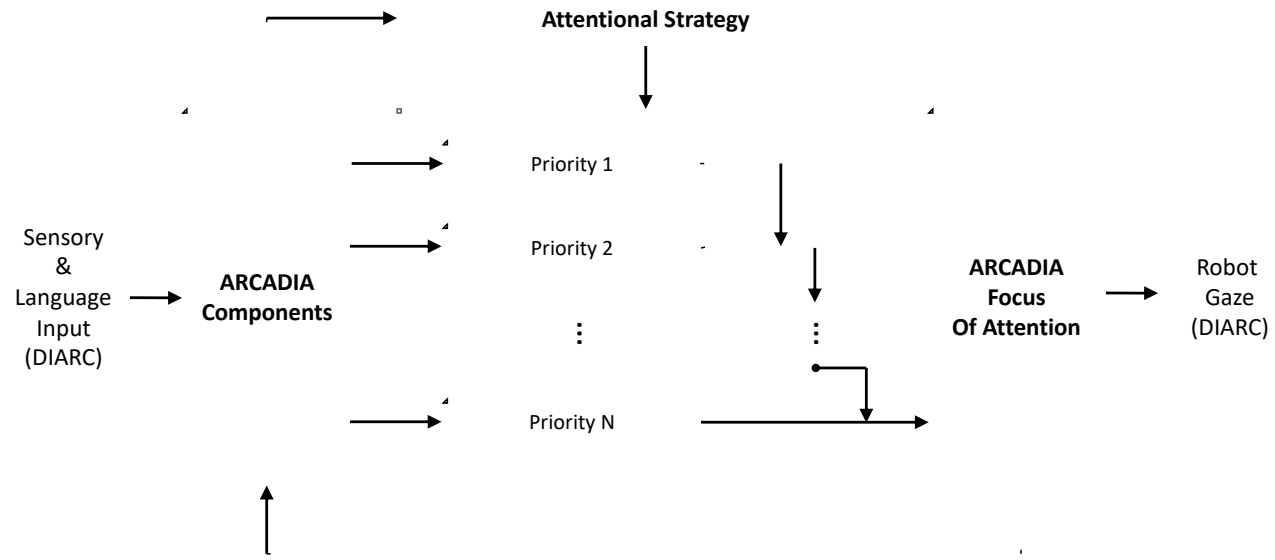

Fig. 2. Diagram depicting how the ARCADIA attentional strategy affects gaze behavior. Sensory, dialogue history, and robot state information are communicated to ARCADIA from DIARC. ARCADIA components then package and process this information into ARCADIA elements. The attentional strategy then selects one of these elements to be the focus of attention, according to the priorities encoded in the strategy. The current focus of attention may affect how information is processed by ARCADIA components, and the attentional strategy can be changed as a result of new information. The focus of attention determines where, if at all, the robot should shift its gaze.

switch to a different strategy that prioritizes online training (i.e., $C \succ ...$). It is not hard to imagine situations that might warrant each different possible prioritization strategy.

ARCADIA's attentional strategies operate in a similar manner, except instead of prioritizing office tasks, it prioritizes pieces of information. Figure 2 depicts how an attentional strategy determines gaze behavior. On each computational cycle, ARCADIA selects the highest priority piece of information available in its information buffer (accessible content) to be the focus of attention. Also, like in the example, ARCADIA's attentional strategies are not necessarily fixed, but can be switched and changed depending on its current cognitive task.

A. Gaze Behaviors

Attentional strategies prioritize ARCADIA's focus on different information elements, which informs DIARC as to where to orient the robot's gaze. We describe three simple strategies that result in different gaze behaviors, below.

1) Inhibition of Return: When we survey our environment, we usually do not fixate on a single object indefinitely. The behavioral bias of avoiding focus on previously or currently focused objects is called *inhibition of return* [25]. As it is a common phenomenon in human visual perception, it has been implemented in prior ARCADIA cognitive models [13]. In the context of our robotic gaze mechanism, inhibition of return is used to prioritize focus on objects not currently within the center of the robot's field of view. If multiple objects exist outside the current center of the field of view, then a random object is selected. This results in the behavior of shifting gaze from object to object.

2) Fixation on Groups: In our daily lives, we sometimes do not fixate on individual objects, but rather on groups of objects. For instance, when enumerating a collection of objects, we often shift eye gaze among salient subgroups of objects, rather than individuals [47]. To enable similar group fixation behavior, we implemented an ARCADIA component that calculated the centroid of all detected 3D objects in the field of view, generating a new 3D point representing the center of this collection. Prioritizing this group centroid enables shifting gaze toward the middle of a group of objects. For example, if two items are detected, gaze would be reoriented toward the center point between the two objects.

3) Fixation on Linguistically Cued Objects: People often shift their gaze among items referred to or possibly referred to in dialogue [44]. Each 3D object representation (`ade-memory-object`) contains a set of visual description predicates associated with the detectors in the DIARC vision system responsible for detecting the object instance (e.g., cup detector - $cup(X)$, red detector - $red(X)$, etc.). The `linguistic-cue` information from DIARC's dialogue history provides a set of visual description predicates information from the most recent utterance. If a `linguistic-cue` element exists, each 3D object can be ranked by similarity to the visual description predicates from this element. For example, if two items are detected, one green and one yellow, and the dialogue system hears "the green cup", then the `ade-memory-object` associated with the green object would be selected as the focus of attention. Gaze is thus directed at objects based on possible relevance to the current dialogue.

B. Composing Gaze Behaviors

Attentional strategies can be written to include different sets of behaviors with different levels of priority (see Figure 2), allowing for the emergence of different, more complex gaze behaviors. For example, a strategy that includes both linguistic cuing and inhibition of return, with linguistically cued objects prioritized, would result in the robot shifting gaze between objects until one of the objects (or a similar looking object) was mentioned in dialogue. Once the linguistic cue was no longer present, either through subsequent, non-relevant dialogue or a timeout, the robot would return to randomly shifting gaze between objects.

C. Early Demonstration

As an initial proof-of-concept demonstration of the integration of ARCADIA/DIARC, we filmed a PR2 exhibiting gaze shifting behavior resulting from a pure inhibition of return attentional strategy. Gaze behavior generated by this simple attentional strategy is depicted in a video found in the supplementary materials[1]. During the beginning of the demonstration, the PR2 shifts its gaze between the two detected objects. As the PR2 centers its gaze on one object, the attentional strategy inhibits its corresponding object representation and selects the non-centered object, and the PR2 shifts its gaze after a short delay. In the middle of the video, a third object is introduced, and the PR2 shifts its gaze to this new object. The robot then subsequently shifts its gaze among the three objects. To demonstrate multiple gaze behaviors working in conjunction with one another, we tested the DIARC/ARCADIA integration in an interactive object learning task. This domain is presented in the following section.

IV. Demonstration: Interactive Object Learning

Consider the following scenario illustrated in Figure 3. A Fetch robot is situated behind a table with two objects: a red cube and a blue cube. The robot engages in the following dialogue interaction with a human teacher:

Teacher: Do you see the *red* object? (1)
Robot: Yes. (2)
Teacher: Do you see the *blue* object? (3)
Robot: Yes. (4)
Teacher: The *red* object is a *glorp*. (5)
Robot: Okay. (6)
Teacher: The *blue* object is a *blicket*. (7)
Robot: Okay. (8)
...
Teacher: Point to the *blicket*. (9)

After each utterance from the human teacher, the robot could engage in one of three possible behaviors, before verbally responding:

1) The robot does not shift its gaze
2) The robot shifts its gaze to the referenced object (e.g., robot looks at the blue object while responding to line 7)
3) The robot shifts its gaze to the opposite, unreferenced object (e.g., robot looks at the red object while responding to line 7)

[1]Videos also found at the following OSF page: https://osf.io/k8pj4/

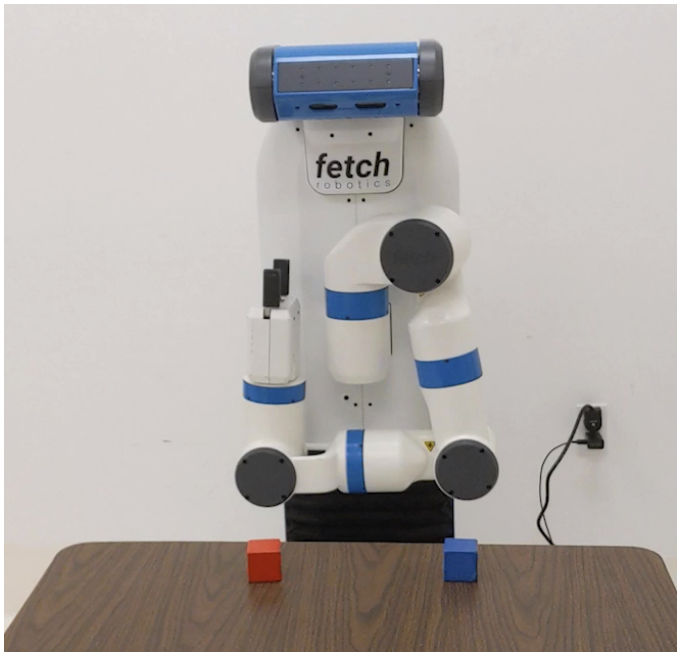

Fig. 3. A tabletop interactive object learning scenario.

We denote behavior 1 as *neutral* gaze behavior, while we denote behaviors 2 and 3 as *adaptive* and *maladaptive* gaze behavior, respectively. Below, we describe what the perceptual and language mechanisms in DIARC and the attentional mechanisms in ARCADIA are doing during this interaction, and how they can be configured to produce behavior corresponding to each gaze behavior.

A. Neutral Gaze

Neutral gaze behavior is the baseline behavior. Using DIARC's natural language, visual perception, and object learning capabilities without the ARCADIA integration enables interactive, dialogue-driven object learning. However, no gaze shifting behavior is present. A video of the interaction with neutral gaze behavior can be found in the supplementary materials.

In the video, the human teacher first exchanges greetings with the Fetch robot. The interaction then proceeds as described previously. First, the teacher asks whether the robot sees the red object (Line 1). If red object detection was not already activated, this triggers DIARC's vision system to search for red objects on the tabletop. The robot finds the red block and replies affirmatively (Line 2). A similar interaction occurs with the blue block (Lines 3-4). Next, the teacher states that the red object is a "glorp," which engages DIARC's object learning mechanisms, associating the with the label "glorp."

(Lines 5-6). The same interaction occurs with regard to the blue block, being assigned the name "blicket" (Lines 7-8). The robot, having correctly learned the labels to each object, points at the blue or red object depending on whether it is instructed to point at the "blicket" or "glorp," respectively.

B. Adaptive Gaze

Video of the learning interaction with adaptive gaze behavior can be found in the supplementary materials. The dialogue and visual learning components of the interaction, enabled by DIARC, are the same as in the neutral gaze section described above. However, now DIARC is also communicating with ARCADIA, operating with an attentional strategy that prioritizes: (1) linguistically cued objects; (2) group centroid fixation; (3) inhibition of return. In the video, the human teacher exchanges greetings with the Fetch robot, as before. Because there is no visual descriptor content in these utterances, there is no linguistic cue that is sent to ARCADIA. If DIARC's object detector was already activated, two 3D object representations corresponding to both objects would be forwarded to ARCADIA's accessible content buffer. This would trigger focus toward the point between the two objects (being the next available element in accessible content). If object detection was not activated, there would be no 3D object representations in ARCADIA's accessible content to select for focus. In either case, the robot's gaze remains in a neutral position looking between both objects.

When the teacher asks whether the robot sees the red object (Line 1), this triggers the activation of a red object detector, which provides additional visual descriptor information to the 3D object representation associated with the red object, and produces a linguistic cue $(red(X), object(X))$ that is forwarded to ARCADIA. The attentional strategy would then prioritize the object representation that best matches the linguistic cue, shifting the robot's gaze to the red object as the robot replies (Line 2). After a few seconds, the linguistic cue decays. A similar chain of events occurs with Lines 3-4, except now the linguistic cue would now prioritize the 3D object representation corresponding to the blue object. As the linguistic cue decays, the robot's gaze returns to the neutral point between the two detected objects. This gaze behavior continues in a similar manner for the rest of the dialogue. [2]

C. Predictions

As mechanisms that enable teaching robots through natural language interactions are developed [41], [42], questions arise regarding how robotic learners are perceived by human teachers. Recent work has examined how gaze behavior affects judgments of the robot's attention to learning and confidence [4]. However, how gaze behavior affects whether human interaction partners are confident in whether or not a robot learner has successfully learned the intended information has yet to be examined. In this regard, the difference between the robot's behavior in the adaptive vs. neutral gaze conditions is stark. Intuitively, people should have more confidence that the robot has learned each object correctly in the adaptive condition compared to the other conditions. Additionally, it seems that maladaptive robot behavior should be associated with the opposite belief, that the robot has learned the objects incorrectly. This leads to the following predictions:

P1: People will more strongly predict that the robot will point to the correct object in the adaptive condition than in the neutral or maladaptive condition.

P2: People will more strongly predict that the robot will point to the incorrect object in the maladaptive condition than in the neutral or adaptive condition.

Additionally, we would predict that people should find the robot in both the adaptive and maladaptive gaze conditions less unresponsive than the robot in the neutral gaze condition. To test these predictions, we conducted an initial evaluation of the system in a crowd-sourced study.

V. Pilot Evaluation

We conducted our study online, using videos that captured the interactions described above. While the original demonstration videos included the robot pointing at learned objects, videos used in the evaluation were edited to stop before the robot shifted its gaze in response to the pointing command and began its pointing action.

A. Participants

A total of 150 U.S.-based participants who were fluent in English and had a high approval rating on Prolific.co were recruited through the platform and participated in the study. Of these, 143 completed all of our measures and passed our attention checks and are included in the analyses below. Participants were between 18 and 54 years-old (M=25.52, SD=7.04 years). The gender distribution for the sample was: male 23.08%, female 74.12%, non-binary 2.8%.

B. Procedure

We used a between-participant design, where participants were divided into three main conditions, depending on the gaze behavior they saw: neutral ($n = 48$), adaptive ($n = 48$) or maladaptive ($n = 47$). Through random assignment, approximately half of the participants in each condition saw the human teacher ask the robot to point to one object, and half to the other object. The videos were filmed from the vantage point of the human teacher and showed a front view of the robot that was positioned behind a table. The two objects were placed on the table in front of the robot (see Fig. 3). The human was not visible in the frame, but his voice could be heard. The videos also contained subtitles of the dialogue between the human and the robot.

After watching the video, participants were asked a series of questions. A static photo of the robot and the two objects

[2]Inverting the linguistic cue similarity calculation produces behavior consistent with maladaptive gaze.

labeled A (red object) and B (blue object) remained on the screen, above the questions. First, we verified that the participants were able to identify in the photo the object that the robot was asked to point to, by matching it to the A or B labels. We used this as an attention check and excluded from analyses participants that were not able to answer correctly. To capture participants predictions of whether the robot will point, and if so whether it will point correctly they were asked the following questions: a) "Use the slider to indicate how likely it is that the robot will point to an object from 0% - Definitely will not point to 100% - Definitely will point." (Slider range: 0-100, increments of 1). b) "If the robot points to an object, use the slider to indicate which object the robot will point to.", with answer anchors: Definitely A (Left), Equally likely A or B (Middle), and Definitely B (Right; Slider range: 0-100, increments of 1). As an additional check, we then asked participants about the object that the robot was not asked to point to, prompting them to indicate how likely it was that the robot had learned to match it to its label. Finally we asked participants to indicate their agreement or disagreement to four statements inspired by [36]: 1) The robot is unresponsive. 2) The robot will malfunction. 3) The robot meets the needs of the task. 4) The robot will perform exactly as instructed. The ratings were indicated by using a slider from 0 - Strongly disagree to 100 - Strongly agree.

C. Results

First, we calculated for each participant the predicted likelihood that the robot will point at the correct object by multiplying the percent chance that the robot will point (question a) with the percent chance that it will point correctly (question b). Shapiro-Wilk tests showed that the data were not normally distributed in either the neutral, $W = .906, p < .001$, adaptive, $W = .912, p = .001$ or maladaptive gaze conditions, $W = .889, p < .001$. We thus proceeded with performing the Kruskal-Wallis non-parametric test to assess whether there was an effect of condition on people's predicted likelihood that the robot will point at the correct object, and found a significant effect , $\chi^2(2) = 6.173, p = 0.046$. To compare the individual conditions, we performed non-parametric pairwise multiple comparisons using Dunn's test. We found that participants in the maladaptive gaze condition predicted a significantly lower chance that the robot will point correctly than participants in the adaptive gaze $p = .016$ and neutral gaze conditions $p = .015$. However, we found no differences between the adaptive gaze and neutral gaze conditions $p = .494$. As an additional confirmation, and to rule out the possibility that people might have believed that the robot's motors were faulty but not its ability to learn, we checked for differences between the maladaptive gaze condition and the other two conditions with regards to the predicted likelihood that the robot had learned to identify the object it was not asked to point to. Dunn's tests revealed a significant difference between the maladaptive gaze and both neutral ($p = 0.033$) and adaptive gaze conditions ($p = 0.044$). We also found significant differences between conditions in participants' agreement with the statement that the robot was unresponsive $\chi^2(2) = 7.974, p = .019$. Dunn's test pairwise comparisons revealed that participants in the neutral gaze condition agreed with this statement more than participants in both the adaptive ($p = .012$) and maladaptive gaze conditions ($p = .004$). For the other statements, we found no significant differences in agreement between conditions.

D. Discussion

The results supported our P2 prediction, that participants would be more likely to predict that the robot in the maladaptive condition would point to the incorrect object. This demonstrates that the gaze behavior generated by the system did have an effect on human perception of the robot. However, the results did not fully support the P1 prediction, finding no significant difference between the adaptive and neutral gaze conditions. Examining the distribution of responses in each condition (see Figure 4) shows that there is a strong bias toward participants responding that the robot would definitely point to the correct object. Even in the maladaptive condition, where intuitively the median response should be skewed toward predicting incorrect pointing, participants were likely to respond that the robot was going to point correctly. Because of this bias, a more subtle difference between neutral and adaptive gaze behavior may not have been detected with the current experimental setup. We hypothesize that this bias could be due to a few possible explanations. First, the objects on the table were relatively close together, producing more subtle shifts in gaze, which may have been harder for participants to perceive and make judgments about. Second, the robot's head shape and the filming angle may have also contributed to difficulty in making judgments about the robot's field of view and gaze direction. Third, the consistent and sensible dialogue responses by the robot may also contribute an assumption of a high-level of robot competency.

Regardless, we did find that the robot in the neutral gaze condition was found to be significantly less responsive than the robots in the adaptive and maladaptive condition. This finding suggests that the gaze behavior generated by the integrated system does provide some helpful feedback to human interactants. Further refinements of our evaluation will be necessary to tease out whether the lack of significant difference in the pointing prediction question was simply an artifact of our study implementation. As alluded to above, these refinements could include stimuli where the objects are spread further apart, necessitating larger and more dramatic shifts in gaze, and additional conditions where only non-verbal feedback is given.

VI. General Discussion

Successful human-robot interaction depends on a variety of factors, including whether people infer the correct beliefs about what these robotic agents have perceived about their environment [45]. Cues such as eye gaze enable people to make such inferences during social interactions [20]. Therefore, studying how gaze behavior can be used by robots to facilitate interactions is an active area of research [2], [3].

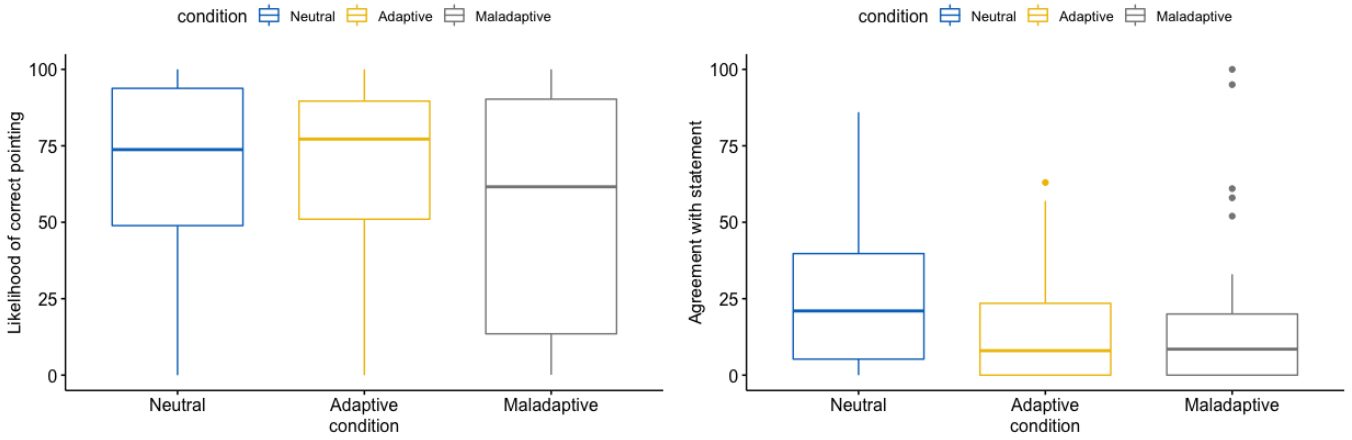

Fig. 4. Box plots representing participants' perceptions that the robot would point to the correct object (left) and that the "robot was unresponsive" (right).

Throughout the years, many approaches to enabling dynamic gaze behavior in robots have been developed (see [2] for a survey). These include approaches with data-driven [1], heuristic [5], [51], and biologically- and cognitively-inspired aspects [12], [24], [46]. One of the challenges the field faces is the diversity of robotic platforms and interactive tasks gaze behavior could be involved in, often making direct comparisons between proposed approaches difficult. Furthermore, attention itself is a multifaceted phenomenon, which is realized differently among cognitive architectures with attentional components (see [26] for a comprehensive review and discussion).

To enable helpful, dynamic gaze behavior in robotic agents, we have proposed and demonstrated the integration of the ARCADIA attention-driven cognitive framework and the DIARC cognitive robotic architecture. The DIARC architecture provides general mechanisms that abstracted over different robotic platforms, while the ARCADIA framework enables abstraction of gaze behavior over different tasks. This integration was demonstrated in different scenarios on multiple robotic platforms. Our initial proof-of-concept demonstration involved a PR2 robot dynamically shifting its gaze between objects in a simple tabletop environment. Our second demonstration involved a Fetch robot dynamically shifting its gaze between objects in a dialogue-driven object learning scenario.

To evaluate the effectiveness of this gaze behavior in helping a human observer feel that the robot learned correctly and was responsive, we conducted a crowdsourced study using videos of the robot operating with and without the integrated system. While the results were not consistent with our prediction regarding increased perceptions of learning success in the adaptive condition over the neutral control condition, the response distribution suggests possible limitations to our current evaluation method. Also, the data did indicate a significant effect of dynamic gaze: maladaptive gaze increased perceptions of incorrect learning and adaptive and maladaptive gaze decreased perceptions of robot unresponsiveness.

In some regards, the attentional strategies enable development of gaze behaviors in a manner akin to a subsumption architecture [7]. In a subsumption architecture, simple behaviors are composed in a hierarchical manner to enable more emergent, complex behaviors. However, unlike in a classical subsumption architecture approach, the ARCADIA/DIARC integration does not eschew symbolic representation. Rather, information elements in both DIARC and ARCADIA are designed to be representationally agnostic, allowing for information processing that can be influenced by both lower-level sensory data and higher-level cognitive processing. Other recent work in generating dynamic gaze behavior for robots also relies on this hierarchical layering of gaze behaviors, explicitly referencing subsumption architectures [33]. Though not demonstrated in the context of this paper, one point of contrast to this approach is that ARCADIA does have the ability to change attentional strategies based on task context, which enables top-down configuration of the hierarchy of gaze behaviors.

As [26] point out, ARCADIA's architecture (specifically its accessible content) resembles Global Workspace Theory [8], a resemblance which is also shared by the ASMO architecture [30], [31]. Like the ARCADIA/DIARC integration, ASMO is an architecture designed to both have an explicit attentional mechanism and ability to control robotic platforms [30]. In contrast to ARCADIA, ASMO's attentional mechanism does not select a unitary focus of attention. Rather, ASMO's attentional mechanism is used to mediate between multiple possible resource (e.g., robotic effectors) requests. Thus, it is possible for ASMO to focus on multiple activities at once, provided these activities do not share resources [30]. Another biologically inspired approach toward robotic gaze control is the STAR architecture, which has recently been used to enable a simulated robot to engage in visual search tasks [34]. While ASMO and STAR are examples of architectures that bridge bottom-up and top-down attention, they do not currently have natural language components, unlike the presented ARCADIA/DIARC integration.

While the behavior exhibited in our two demonstrations

is relatively simple, the integration that it demonstrates is powerful. Any information available from DIARC components can be given to ARCADIA, processed, and used to potentially influence the focus of attention. There is a considerable amount of flexibility in which of the two architectures is responsible for specific pieces of information processing and decision-making. For instance, ARCADIA can receive information about key features of the robot's environment and state, and use this to compute where to look. Other configurations could use ARCADIA not only to decide where to look, but to model more human-like and attention-dependent cognitive processes. In these cases, the DIARC wrapper component could potentially initiate different postures and gaze behaviors reflective of attention toward internal representations rather than external objects or points in space.

Generating helpful and informative gaze behavior during dialogue interactions is a common application for models of gaze behavior [11], [28], [46]. As demonstrated, the ARCADIA/DIARC integration can direct gaze toward objects mentioned in dialogue, though future extensions of this work are needed to bias and direct gaze toward speaking agents. For example, sound localization and speech onset information can be used to know when and where to direct the robot's gaze when an interaction partner is speaking. Furthermore, dialogue state information from DIARC's dialogue system could be used to generate anticipatory gaze shifts [35] based on the expected next speaker (see [48] for an example of top-down biasing in DIARC based on dialogue state information). We would predict that a robot exhibiting both reactive and anticipatory gaze shifts in dialogue interactions would be viewed as a more natural interaction partner by human interactants.

Another direction of future work is using the ARCADIA/DIARC integration to enable active perception [9], using the attentional mechanism to cue gaze and robot orientation toward areas outside the current field-of-view to gather more information about the environment. Additionally, using ARCADIA's capacity to change attentional strategies during task-switching would enable context-dependent gaze behavior [15]. For instance, in a collaborative scenario during which a robot must both manipulate items in an environment and communicate with a human interaction partner, subtasks that involve critical manipulation actions could deprioritize gaze toward the interaction partner, whereas other subtasks could involve switching gaze between the human partner and the task items.

Finally, different ways to integrate ARCADIA and DIARC can be explored. For example, each DIARC component could potentially have a separate instance of ARCADIA to provide an attentional processing mechanism dedicated to the function of that specific component. In this paper, we have presented the foundation that enables these possibilities.

Acknowledgments

This work was supported by awards N0001421WX01321 and N0001420WX00908 by the Office of Naval Research. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

[1] H. Admoni and B. Scassellati, "Data-driven model of nonverbal behavior for socially assistive human-robot interactions," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 196–199.

[2] ——, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

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

[4] P. Aliasghari, M. Ghafurian, C. L. Nehaniv, and K. Dautenhahn, "Effects of gaze and arm motion kinesics on a humanoid's perceived confidence, eagerness to learn, and attention to the task in a teaching scenario," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 197–206.

[5] P. Aliasghari, A. Taheri, A. Meghdari, and E. Maghsoodi, "Implementing a gaze control system on a social robot in multi-person interactions," *SN Applied Sciences*, vol. 2, pp. 1–13, 2020.

[6] J. R. Anderson, "Act: A simple theory of complex cognition." *American psychologist*, vol. 51, no. 4, p. 355, 1996.

[7] R. C. Arkin, R. C. Arkin *et al.*, *Behavior-based robotics*. MIT press, 1998.

[8] B. J. Baars, "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience," *Progress in brain research*, vol. 150, pp. 45–53, 2005.

[9] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[10] P. Bello, A. M. Lovett, G. Briggs, and K. O'Neill, "An attention-driven computational model of human causal reasoning." in *CogSci*, 2018.

[11] 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*, 2010, pp. 1–8.

[12] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *rn*, vol. 255, no. 3, 1999.

[13] W. Bridewell and P. Bello, "Inattentional blindness in a coupled perceptual-cognitive system." in *CogSci*, 2016.

[14] ——, "A theory of attention for cognitive systems," *Advances in Cognitive Systems*, vol. 4, no. 1, pp. 1–16, 2016.

[15] W. Bridewell, C. Wasylyshyn, and P. F. Bello, "Towards an attention-driven model of task switching," *Advances in Cognitive Systems*, pp. 85–100, 2018.

[16] G. Briggs, W. Bridewell, and P. Bello, "A computational model of the role of attention in subitizing and enumeration." in *CogSci*, 2017.

[17] G. Briggs, T. Williams, and M. Scheutz, "Enabling robots to understand indirect speech acts in task-based interactions," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 64–94, 2017.

[18] D. Buckingham, D. Kasenberg, and M. Scheutz, "Simultaneous representation of knowledge and belief for epistemic planning with belief revision," in *Proceedings of KR*, 2020.

[19] M. M. de Graaf, B. F. Malle, A. Dragan, and T. Ziemke, "Explainable robotic systems," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 387–388.

[20] N. J. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *Neuroscience & biobehavioral reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[21] T. Frasca, Z. Han, J. Allspaw, H. Yanco, and M. Scheutz, "Going cognitive: A demonstration of the utility of task-general cognitive architectures for adaptive robotic task performance," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8110–8116.

[22] F. Gervits, D. Thurston, R. Thielstrom, T. Fong, Q. Pham, and M. Scheutz, "Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models," in *Proceedings of AAMAS*, 2020.

[23] M. Hartmann and M. H. Fischer, "Exploring the numerical mind by eye-tracking: a special issue," 2016.

[24] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
[25] R. M. Klein, "Inhibition of return," *Trends in cognitive sciences*, vol. 4, no. 4, pp. 138–147, 2000.
[26] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: core cognitive abilities and practical applications," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 17–94, 2020.
[27] J. E. Laird, *The Soar cognitive architecture*. MIT press, 2019.
[28] J. Lee, S. Marsella, D. Traum, J. Gratch, and B. Lance, "The rickel gaze model: A window on the mind of a virtual human," in *International workshop on intelligent virtual agents*. Springer, 2007, pp. 296–303.
[29] A. Lovett, W. Bridewell, and P. Bello, "Selection enables enhancement: An integrated model of object tracking," *Journal of vision*, vol. 19, no. 14, pp. 23–23, 2019.
[30] R. Novianto, "Flexible attention-based cognitive architecture for robots," Ph.D. dissertation, 2014.
[31] R. Novianto, B. Johnston, and M.-A. Williams, "Attention in the asmo cognitive architecture," in *Biologically Inspired Cognitive Architectures 2010*. IOS Press, 2010, pp. 98–105.
[32] B. Oosterveld, L. Brusatin, and M. Scheutz, "Two bots, one brain: Component sharing in cognitive robotic architectures," in *Proceedings of HRI, Video Contest*, 2017.
[33] M. K. Pan, S. Choi, J. Kennedy, K. McIntosh, D. C. Zamora, G. Niemeyer, J. Kim, A. Wieland, and D. Christensen, "Realistic and interactive robot gaze," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 072–11 078.
[34] A. Rasouli, P. Lanillos, G. Cheng, and J. K. Tsotsos, "Attention-based active visual search for mobile robots," *Autonomous Robots*, vol. 44, no. 2, pp. 131–146, 2020.
[35] C. Riest, A. B. Jorschick, and J. P. de Ruiter, "Anticipation in turn-taking: mechanisms and information sources," *Frontiers in psychology*, vol. 6, p. 89, 2015.
[36] K. Schaefer, "The perception and measurement of human-robot trust," 2013.
[37] P. Schermerhorn and M. Scheutz, "Natural language interactions in distributed networks of smart devices," *International Journal of Semantic Computing*, vol. 2, no. 4, pp. 503–524, 2008.
[38] M. Scheutz, "ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures," *Applied Artificial Intelligence*, vol. 20, no. 4-5, 2006.
[48] R. Veale, G. Briggs, and M. Scheutz, "Linking cognitive tokens to biological signals: Dialogue context improves neural speech recognizer performance," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, no. 35, 2013.
[39] M. Scheutz, S. DeLoach, and J. Adams, "A framework for developing and using shared mental models in human-agent teams," *Journal of Cognitive Engineering and Decision Making*, vol. 11, no. 3, pp. 203–224, 2017.
[40] M. Scheutz, J. Harris, and P. Schermerhorn, "Systematic integration of cognitive and robotic architectures," *Advances in Cognitive Systems*, pp. 277–296, 2013.
[41] M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt, "Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture," in *Proceedings of AAMAS*, 2017.
[42] ——, "Recursive spoken instruction-based one-shot object and action learning," in *Proceedings of IJCAI*, 2018, pp. 5354–5358.
[43] M. Scheutz, T. Williams, E. Krause, B. Oosterveld, V. Sarathy, and T. Frasca, "An overview of the distributed integrated cognition affect and reflection diarc architecture," *Cognitive architectures*, pp. 165–193, 2019.
[44] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, no. 5217, pp. 1632–1634, 1995.
[45] S. Thellman and T. Ziemke, "The perceptual belief problem: Why explainability is a tough challenge in social robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–15, 2021.
[46] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 2008, pp. 201–208.
[47] M. P. Van Oeffelen and P. G. Vos, "Enumeration of dots: An eye movement analysis," *Memory & Cognition*, vol. 12, no. 6, pp. 607–612, 1984.
[49] T. Williams, G. Briggs, B. Oosterveld, and M. Scheutz, "Going beyond command-based instructions: Extending robotic natural language interaction capabilities," in *Proceedings of AAAI*, 2015.
[50] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*. Springer, 1967, pp. 171–211.
[51] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto, "Responsive robot gaze to interaction partner." in *Robotics: Science and systems*. Citeseer, 2006, pp. 37–43.
[52] C. Yu, P. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, p. 13, 2012.