Now, Over Here: Leveraging Extended Attentional Capabilities in Human-Robot Interaction

Xiang Zhi Tan[1,2], Sean Andrist[1], Dan Bohus[1], Eric Horvitz[1]
[1]Microsoft Research, Redmond, WA, USA [2]Carnegie Mellon University, Pittsburgh, PA, USA
zhi.tan@ri.cmu.edu,{sandrist,dbohus,horvitz}@microsoft.com

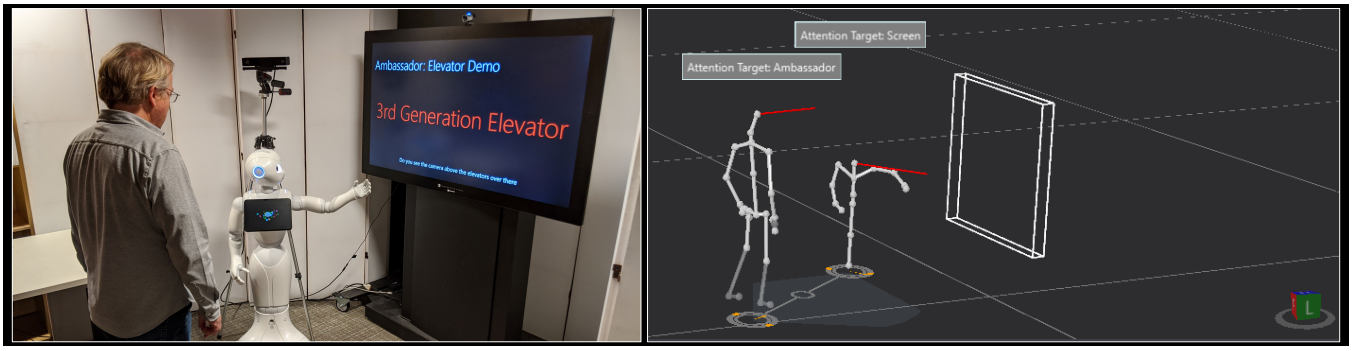

Figure 1: Left: The Ambassador, a research prototype for investigating challenges with tracking, reasoning about, and managing attention. Right: 3D visualization of the situated context, including directions and targets of attention.

ABSTRACT

Competent collaboration between robots and people in the open world requires sensing and reasoning about transitions of people's attention to the robots themselves, as well as to other people and objects in the environment. We present challenges and opportunities with designing extended attentional capabilities for interactive systems, including the need to track, reason about, and manage the attentional foci of all actors. We describe work in progress to leverage such attentional capabilities for interaction management with a prototype situated robotic system.

CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools*; *Systems and tools for interaction design.*

KEYWORDS

Visual Attention; Interaction Management; Situated Interaction

ACM Reference Format:
Xiang Zhi Tan, Sean Andrist, Dan Bohus, and Eric Horvitz. 2020. Now, Over Here: Leveraging Extended Attentional Capabilities in Human-Robot Interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378363

HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378363

1 INTRODUCTION

To date, human-robot interaction (HRI) has focused largely on the dyadic engagement of people and robots in controlled settings. As social robotic systems begin to interact in open environments, the scope of interactions must grow to consider details of the surrounding context, including how attention of the robot and other parties is distributed across all actors and relevant objects in the environment. For example, robots may use external screens to enhance the interaction [4], interact with other robots [15], or refer to external objects [3, 11]. In such cases, the flow of interaction must consider not only attention towards the robot but also towards other entities and objects. Our behaviors as humans are continually influenced by the attentional foci of others. For example, a speaker may hesitate or restart if an interlocutor appears to be distracted [5] and even employ different non-verbal behaviors to gain or redirect joint attention [8]. A speaker may also take others' spatial perspectives into account when referring to objects and locations [13].

Prior research has shown how robotic technologies can leverage the visual attention of users to gauge understanding and intent [2, 6, 9, 10] and use their own gaze to establish joint attention to task-relevant objects [1, 2, 7, 12]. Yu et al. [16] described a speech production model that coordinated with the listener's attention by inserting pauses, re-starts and interjections. Stefanov et al. [14] described a model to generate an agent's visual attention that matched synthesized speech in a multi-party dialog scenario. However, there has been little work to date on human attention as a fundamental pillar of interaction management. In many cases, attention is treated as a secondary construct aimed at supplementing dialogue.

Creating effective open-world interactions with social robots will require more sophisticated inference and actions aimed at leveraging and managing attention. Attentional considerations can be as

important as spoken utterances and task actions in driving interactions. To leverage attentional considerations effectively, robots need to track all relevant objects in the scene and consider the potential importance of different objects to the interaction. They must also consider each user's gaze and visual attention at a semantic level over time while jointly reasoning about intentions, turn-taking, engagement, and grounding. Furthermore, robots must decide and plan how to convey their own visual attention while managing joint attention with others.

In this report, we outline several key challenges and opportunities in developing extended attentional capabilities for human-robot interactive systems. We then describe the initial version of a prototype robotic system that considers the attention of users in controlling the flow of interaction (Figure 1).

2 KEY CHALLENGES

Seamless participation in physically situated interaction hinges on extended attentional capabilities to (1) accurately *track* users' attention, (2) *reason* about attention with respect to engagement, turn-taking, and grounding, and (3) jointly *manage* attention for interaction planning with users.

Tracking Direction and Target of Attention: A prerequisite for enabling higher-level attentional behaviors is the ability to track both the continuous 3D directions in which users are looking and to identify the entities (actors or objects) that users are attending to. Inferences about the direction of attention are typically based on eye gaze or head pose estimation [2]; both remain challenging vision problems in open-world settings. To accurately determine the target of a user's attention, systems need to track relevant entities in the environment, reason about them and their relationships in the context of the task, and harness this knowledge in conjunction with higher-level expectations about the user's attention.

Attention-Based Reasoning: Visual attention can reveal specific higher-level intentions and inform communicative processes. For instance, loss of attention for an extended period of time may signal loss of engagement, gazing away from the listener at the end of a turn may signal the intention to keep the conversational floor, looking (or not looking) at a relevant object may indicate lack of grounding, etc. Models for engagement, turn-taking, and grounding therefore need to deeply leverage information about the users' attention, together with broader contextual information.

Management and Control: Attention-based inferences can also shape decisions and streamline the collaboration at the task level. For instance, attention on a certain relevant object may reveal users' intentions as part of the collaboration, and the system may adjust its strategy in response. Attention itself can be viewed as a resource that should be collaboratively managed during interactions. The robot should be able to not only observe, but also to shape, guide, and re-focus users' attention to certain entities of interest. Various strategies can be employed, from subtle gaze redirection to a combination of speech, gaze, and hand gestures. The ideal choice of behaviors depends on the state of the interaction, the intended target, as well as the physical surrounding context.

When managing attention, the system needs to be able to take perspective and reason about potential optical occlusions and barriers. Consider for instance the case of a robot gesturing towards a distant object, but where the line of sight is blocked by other people, such as passersby or other interaction participants. The best way to address such a situation depends on the task at hand (e.g., does an alternative plan exist?) and on whether people can or will move (e.g., just wait a moment or ask them to shift?). Developing general policies for managing joint attention, from first principles, is an open research area. Solutions will contribute to more seamless collaborations between humans and robots.

The challenges of tracking, reasoning about, and managing attention are best addressed together as they are tightly coupled and operate in synergy. The multiparty, dynamic nature of open-world environments further amplifies these challenges and opportunities.

3 THE AMBASSADOR SYSTEM

We have started to explore these challenges in the context of the *Ambassador*, a prototype robotic system we are developing to serve as an institutional guide in our building (Figure 1). The system will be able to describe and talk about various research projects at our organization. The physical platform consists of a Softbank Pepper robot, supplemented by external sensors and a large touch-enabled display to present visual information. Central to the application is the ability to present information in a coherent manner while redirecting users' attention across multiple targets, such as the robot itself, the display, and other relevant objects in the environment. Relevant entities in the environment are pre-specified, and the system infers the user's direction and target of attention from tracked body and head orientations.

As a first scenario, the robot demonstrates and describes an intelligent elevator system in its vicinity. To accomplish this task, the robot must be able to direct users' attention to the elevators and to a ceiling camera mounted above them. The robot deploys an array of escalating strategies to direct attention to the elevator camera. It starts with either gaze or gesture alone and eventually escalates to a combination of gesture, speech, and gaze until there is joint attention to the selected target. When describing the elevator system, the robot utilizes the screen to present additional information and expects users' attention to move between the robot and the screen. The robot determines the start time of its own speech based on the users' attention and waits for attention on itself before continuing the interaction. When showing a video on the screen, the system uses a similar escalating strategy to direct attention to the screen. The video is paused if the users attend elsewhere and continues only after regaining attention.

These strategies and behaviors are initial examples from our first prototype. We plan to continue exploring the key challenges for extended attentional capabilities with the Ambassador system. We are working to raise the centrality of attention in HRI, with a focus on the need to develop and leverage machinery for tracking, reasoning about, and managing attention in open-world settings. We believe that these considerations are important in natural human-robot collaborations, and that they are especially critical for robots deployed in the open world.

ACKNOWLEDGMENTS

This work was conducted when the first author was an intern with Microsoft Research AI.

REFERENCES

[1] Henny Admoni, Anca Dragan, Siddhartha S. Srinivasa, and Brian Scassellati. 2014. Deliberate Delays during Robot-to-Human Handovers Improve Compliance with Gaze Communication. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction - HRI '14*. ACM Press, Bielefeld, Germany, 49–56. https://doi.org/10.1145/2559636.2559682

[2] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 2571–2582.

[3] Nils Axelsson and Gabriel Skantze. 2019. Modelling Adaptive Presentations in Human-Robot Interaction Using Behaviour Trees. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Stockholm, Sweden, 345–352.

[4] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3, 21 (2018), eaat5954.

[5] Charles Goodwin. 1980. Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry* 50, 3-4 (1980), 272–302.

[6] Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *The eleventh ACM/IEEE international conference on human robot interaction*. IEEE Press, 83–90.

[7] M. Imai, T. Ono, and H. Ishiguro. 2003. Physical Relation and Expression: Joint Attention for Human-Robot Interaction. *IEEE Transactions on Industrial Electronics* 50, 4 (Aug. 2003), 636–643. https://doi.org/10.1109/TIE.2003.814769

[8] Mardi Kidwell and Don H Zimmerman. 2007. Joint attention as action. *Journal of Pragmatics* 39, 3 (2007), 592–611.

[9] Dimosthenis Kontogiorgos, Andre Pereira, and Joakim Gustafson. 2019. Estimating Uncertainty in Task-Oriented Dialogue. In *2019 International Conference on Multimodal Interaction*. 414–418.

[10] Ulyana Kurylo and Jason R Wilson. 2019. Using Human Eye Gaze Patterns as Indicators of Need for Assistance from a Socially Assistive Robot. In *International Conference on Social Robotics*. Springer, 200–210.

[11] Phoebe Liu, Dylan F Glas, Takayuki Kanda, and Hiroshi Ishiguro. 2018. Learning proactive behavior for interactive social robots. *Autonomous Robots* 42, 5 (2018), 1067–1085.

[12] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zeng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 2014. 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 - HRI '14*. ACM Press, Bielefeld, Germany, 334–341. https://doi.org/10.1145/2559636.2559656

[13] Michael F Schober. 1993. Spatial perspective-taking in conversation. *Cognition* 47, 1 (1993), 1–24.

[14] Kalin Stefanov, Giampiero Salvi, Dimosthenis Kontogiorgos, Hedvig Kjellström, and Jonas Beskow. 2019. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *ACM Transactions on Human-Robot Interaction* 8, 2 (June 2019), 1–21. https://doi.org/10.1145/3323231

[15] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. 2019. From One to Another: How Robot-Robot Interaction Affects Users' Perceptions Following a Transition Between Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 114–122.

[16] Zhou Yu, Dan Bohus, and Eric Horvitz. 2015. Incremental Coordination: Attention-Centric Speech Production in a Physically Situated Conversational Agent. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, 402–406. https://doi.org/10.18653/v1/W15-4652