

Deciding Where to Look: Gaze Allocation for Joint Attention Under Uncertainty in HRI

Ivy Xiao He
Brown University
Providence, USA
xiao_he@brown.edu

Jason Xinyu Liu
Brown University
Providence, USA
xinyu_liu@brown.edu

Stefanie Tellex
Brown University
Providence, USA
stefie10@cs.brown.edu

ABSTRACT

Gaze plays a fundamental role in human–robot interaction but is often treated as a passive signal rather than an explicit decision variable. We reframe robotic gaze as a decision-making problem under uncertainty and propose a comparative framework contrasting explicit decision-theoretic and implicit learning-based gaze strategies. Our goal is to clarify how modeling assumptions influence coordination stability, ambiguity resolution, and recovery behavior in collaborative tasks.

1 INTRODUCTION

Gaze signals are essential for naturalistic human–robot collaboration because they serve as a two-way channel: objectively, gaze enables robots to infer human focus and intention, improving efficiency in tasks such as object search and manipulation; subjectively, gaze behaviors shape how intuitive, engaging, and socially natural robots are perceived to be[1, 15]. When and where a robot directs its gaze fundamentally shapes both what information it can perceive and how it coordinates with humans during collaboration. Prior work in HRI has shown that gaze supports natural coordination and is widely used as an informative cue in robotic systems. However, when gaze is incorporated into robotic decision-making, it is often implicitly treated as a reliable signal. In practice, human gaze is ambiguous and fallible, and gaze alignment does not guarantee task correctness—shared attention can propagate shared errors rather than resolve uncertainty.

This project aims to reframe robot gaze as a decision variable under uncertainty, rather than as a passive signal or fixed communicative behavior. We aim to ask: when should a robot attend to the human versus the workspace, and what should it look at to reduce task ambiguity while maintaining smooth coordination? We focus on non-humanoid robots that use camera orientation or body pose as a proxy for gaze. We propose a comparative framework that contrasts explicit, decision-theoretic gaze strategies with implicit, end-to-end policies that exhibit planning-like behavior. Rather than enforcing algorithmic equivalence, we compare these approaches along shared interaction-level dimensions such as gaze timing, robustness to misleading human cues, recovery behavior, and coordination stability. This work aims to clarify the assumptions underlying different gaze modeling paradigms and to inform the design of robust gaze policies for human–robot collaboration.

2 PROBLEM STATEMENT

We aim to study the question: when and what should a robot look during human–robot collaboration to reduce task ambiguity and support effective coordination under uncertainty?

Imagine a robot assisting a person with an assembly task in which several tools are placed on a table but are out of the person’s reach. The robot can observe where the person is looking, but must still infer which tool the person intends to use next. To do this effectively, the robot must decide when to attend to the human and when to focus on the workspace—for example, whether to look at the person’s face to assess attention or uncertainty, or to inspect candidate tools to confirm its hypothesis. Importantly, human gaze may be ambiguous or misleading: a person may glance at multiple tools, hesitate, or look away while thinking. Blindly following gaze cues can therefore lead to premature or incorrect alignment.

The challenge is to formulate a gaze control policy that decides how to allocate visual attention between the human and the workspace, balancing information gathering with smooth and natural coordination. This leads to the following research questions: This leads to the following research questions:

This leads to the following research questions:

- (1) **When should the robot switch its attention?** Under what uncertainty, timing, or interaction cues should the robot allocate gaze to the human (e.g., to interpret engagement or confusion) versus the workspace?
- (2) **Where should the robot look in the workspace?** Given multiple candidate targets, which objects or regions should the robot inspect to reduce ambiguity and verify its current hypothesis about the human’s intent?
- (3) **How does gaze allocation affect collaboration dynamics?** How do different gaze strategies influence task efficiency, coordination stability (e.g., alignment persistence and jitter), and recovery from misleading or ambiguous human cues over time?

3 RELATED WORK

Current research in gaze focuses on five main areas: perception and intention inference[2, 14], communicative gaze[11], joint attention and coordination models[3, 4], multimodal fusion of gaze[9, 17], and social and psychological effects of gaze[6, 8]. The proposed project aims to bridge joint attention modeling and communicative gaze control by focusing on timing and target selection policies for non-humanoid robots in collaborative tasks. We aim to treat gaze not only as perception or expression, but as a decision variable for coordination. Prior work in human–robot interaction has demonstrated that gaze and joint attention play a central role in coordination, communication, and user experience. However, how robots should reason about gaze under ambiguity—particularly when gaze cues are misleading or incomplete—remains underexplored. In this section, we review relevant work on social gaze, its limitations, and how gaze has been incorporated into robotic decision-making and learning.

Social Gaze in Human–Robot Interaction: A large body of HRI research shows that robot gaze supports coordination, turn-taking, and shared task understanding. Admoni and Scassellati provide a comprehensive review demonstrating that gaze behaviors improve collaboration efficiency and social acceptance across a wide range of tasks [1]. Mutlu and colleagues show that gaze timing and mutual attention cues influence compliance, engagement, and perceived fluency during collaborative tasks such as handovers and instruction following [12, 13]. In collaborative manipulation settings, gaze has also been shown to facilitate grounding and anticipation. Martin-Martin et al. demonstrate that human gaze and motion cues provide predictive signals that can improve object handover and manipulation performance when interpreted correctly [10].

Limits of Gaze: Despite its benefits, gaze is an inherently ambiguous signal in both human perception and human–robot interaction contexts. Human fixation does not always reflect attention or intention: gaze patterns can be influenced by factors such as context, task demands, and perceptual biases [16, 21]. Experimental studies in cognitive science show that perceived gaze direction can bias observers’ inferences about others’ intentions, even when gaze does not directly indicate a goal state [5]. Prior HRI work often assumes correct interpretation of gaze cues, yet many models treat gaze as a stable indicator of intent without explicitly modeling uncertainty or error in gaze interpretation [1, 15]. Joint attention and gaze alignment do not guarantee task correctness: shared attention may propagate errors, particularly in cluttered environments or when human gaze reflects exploration or hesitation rather than communicative intent. These limitations suggest that gaze should be treated as uncertain evidence whose utility depends on context, and that robotic systems should explicitly model gaze unreliability and ambiguity rather than assume gaze fidelity by default.

Gaze in Robotic Decision-Making and Object Search: In robotics, gaze-related cues are often incorporated into object search and manipulation systems as observational inputs. Decision-theoretic approaches to collaborative disambiguation explicitly model uncertainty over human intent using multimodal signals such as language, pointing, head pose, and gaze [18, 19]. However, in most of these systems, gaze is treated as an observation rather than as a controllable sensing action. The robot reasons about what gaze indicates, but not about where it should look to acquire more informative observations. As a result, the timing and target selection of robot gaze are typically fixed or heuristic-driven, rather than optimized as part of the decision process.

Learning Gaze Policies: Implicit and End-to-End Approaches: Recent work in robotics has studied active perception and view selection for efficient object search and task execution, including learning where to look as part of a perception or control policy [7]. These approaches treat gaze or viewpoint selection as a means of optimizing perceptual efficiency, enabling robots to acquire task-relevant information more quickly or reliably. However, gaze in these systems is primarily optimized for perception, and is not considered as a coordination signal in human–robot interaction. As a result, such methods do not analyze how gaze allocation decisions influence joint attention, miscoordination, or recovery from incorrect alignment with a human partner.

More recently, learning-based approaches have explored acquiring gaze and perception behaviors directly from human demonstrations. Vision in Action proposes an end-to-end visuomotor policy that jointly predicts manipulation actions and head movements, allowing robots to learn active perception behaviors implicitly from data [20]. These results demonstrate that gaze-like behaviors and planning-like perception strategies can emerge through learning without explicit modeling.

Despite their effectiveness, end-to-end approaches do not explicitly represent uncertainty or decision trade-offs associated with gaze. Planning-like behavior is encoded implicitly in the learned policy, making it difficult to analyze when gaze helps, when it misleads, and how a robot should recover from incorrect or ambiguous gaze alignment. In contrast, our work focuses on comparing such implicit, learned gaze strategies with explicit, decision-theoretic reasoning, with the goal of understanding how different modeling assumptions shape gaze behavior, coordination dynamics, and failure modes in human–robot collaboration.

4 METHODS & EVALUATION

We study robot gaze as a decision-making problem under uncertainty through a comparative lens that contrasts implicit learned gaze policies and explicit model-based gaze reasoning. Rather than benchmarking performance or identifying a single optimal gaze strategy, our goal is to examine how different modeling assumptions about gaze shape robot behavior and coordination outcomes in HRI.

As concrete instantiations, we consider two representative paradigms as shown in Fig.1. In the explicit paradigm, gaze is treated as a controllable sensing action. The robot maintains an explicit representation of uncertainty over the task and the human’s intent, and selects where to look—such as attending to the human or inspecting the workspace—to acquire informative observations. This approach makes assumptions about gaze reliability and uncertainty explicit, enabling analysis of information-seeking behavior and recovery from incorrect alignment. We instantiate this paradigm using a POMDP-based formulation for gaze allocation.

In contrast, learning-based approaches acquire gaze behavior implicitly from data. Here, gaze decisions are encoded within a policy that maps observations to actions, and planning-like behavior emerges without an explicit representation of belief or information gain. Such policies can exhibit flexible and human-like gaze patterns, but the underlying decision trade-offs are not directly represented. We instantiate this paradigm using a diffusion-policy-based visuomotor model that learns head and attention behaviors from human demonstrations.

These paradigms are not algorithmically aligned; therefore, we do not enforce a one-to-one comparison. Instead, we evaluate their interaction-level consequences under matched task and sensing conditions, focusing on dimensions such as gaze timing, robustness to misleading human cues, recovery behavior, and coordination stability. This evaluation is intended to expose qualitative differences in behavior and failure modes, rather than to determine a universally superior method.

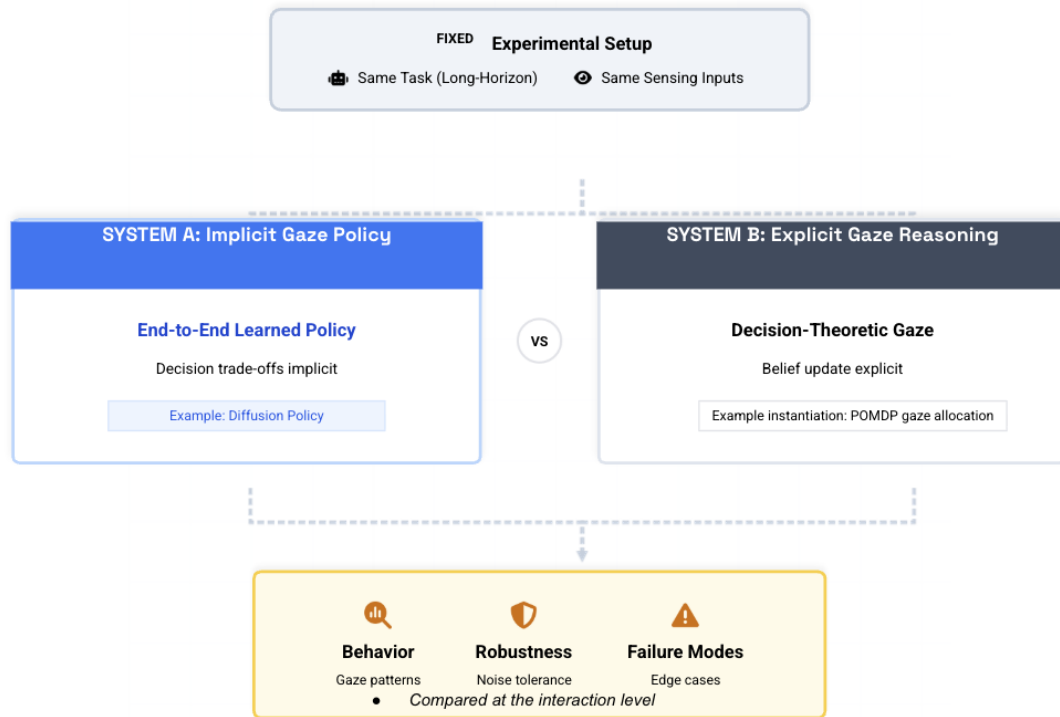


Figure 1: Comparison framework for robot gaze modeling under uncertainty. We fix the task setting and sensing inputs, and contrast two modeling paradigms.

5 EXPECTED CONTRIBUTIONS

This work aims to reframe robot gaze as a decision-making problem under uncertainty in HRI. We distinguish joint attention from task correctness, analyzing how gaze can both facilitate coordination and propagate errors. We propose a comparative framework to study how explicit and implicit gaze modeling assumptions shape robot behavior, coordination dynamics, and failure modes.

REFERENCES

- [1] Henny Admoni and Brian Scassellati. Social eye gaze in human–robot interaction: A review. *Journal of Human–Robot Interaction*, 6(1):25–63, September 2017.
- [2] A. Belardinelli et al. Gaze-based intention estimation: Principles, methodologies, and applications in hri. *ACM Transactions on Human–Robot Interaction*, 13(2): 1–32, 2024. doi: 10.1145/3656376.
- [3] K. Belhassen et al. Addressing joint action challenges in human–robot interaction. *Cognition*, 220:104917, 2022. doi: 10.1016/j.cognition.2021.104917.
- [4] J. Garcia-Martinez et al. Are robots more engaging when they respond to joint attention? *Applied Sciences*, 15(15):8684, 2025. doi: 10.3390/app15158684.
- [5] Matthew Hudson, Chang Hong Liu, and Tjeerd Jellema. Anticipating intentional actions: The effect of eye gaze direction on the judgment of head rotation. *Cognition*, 112(3):423–434, 2009. ISSN 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2009.06.011>. URL <https://www.sciencedirect.com/science/article/pii/S0010027709001541>.
- [6] Magnus Jung et al. Eye contact based engagement prediction for efficient human–robot interaction. *Complex Intelligent Systems*, 11:286, 2025. doi: 10.1007/s40747-025-01902-z.
- [7] Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. *arXiv preprint arXiv:2506.10968*, 2025.
- [8] M. Koller et al. Robotic gaze and human views: A systematic exploration of robotic gaze aversion and its effects on human behaviour and attitudes. *Frontiers in Robotics and AI*, 10:1062714, 2023. doi: 10.3389/frobt.2023.1062714.
- [9] Yuzhi Lai, Shenghai Yuan, Boya Zhang, Benjamin Kiefer, Peizheng Li, and Andreas Zell. Fam-hri: Foundation-model assisted multi-modal human-robot interaction combining gaze and speech. In *arXiv Preprint*, 2025. arXiv:2503.16492.
- [10] Roberto Martin-Martín, Frederik Ebert, Chenxi Wu, Li Fei-Fei, Silvio Savarese, and Jeannette Bohg. Predicting human intent for object handover. In *Proceedings of Robotics: Science and Systems (RSS)*, 2019.
- [11] Chinmaya Mishra and Gabriel Skantze. Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1201–1208, 2022. doi: 10.1109/RO-MAN53752.2022.9900740.
- [12] Ajung Moon, Minhua Zheng, [Daniel M.] Troniak, [Benjamin A.] Blumer, Brian Gleeson, Karon MacLean, [Matthew K.X.J.] Pan, and [Elizabeth A.] Croft. Meet me where i’m gazing: How shared attention gaze affects human-robot handover timing. In *HRI 2014 - Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 334–341, United States of America, 2014. IEEE, Institute of Electrical and Electronics Engineers. ISBN 9781450326582. doi: 10.1145/2559636.2559656. Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2014, HRI 2014.
- [13] Bilge Mutlu, Takayuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human–robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction*, pages 61–68, 2009.
- [14] L. Naendrup-Poell et al. Predictive robot eyes enhance attentional guidance in human–robot interaction. In *Scientific Reports*, 2025. doi: 10.1038/s41598-025-19497-3.
- [15] K. Ruhland, Christopher E. Peters, Sean Andrist, J. Badler, N. Badler, Michael Gleicher, Bilge Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. *Comput. Graph. Forum*, 2015. doi: 10.1111/cgf.12603.
- [16] Lei Shi, Cosmin Copot, and Steve Vanlanduit. What are you looking at? detecting human intention in gaze based human-robot interaction, 2019. URL <https://arxiv.org/abs/1909.07953>.
- [17] Hui Su, Wenjie Li, Shanqing Li, and Xiaoxia Liu. Recent advancements in multi-modal human–robot interaction. *Multimodal Technologies and Interaction*, 7(10): 139, 2023. doi: 10.3390/mti7100139.

- [18] David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. Interpreting Multimodal Referring Expressions in Real Time. In *International Conference on Robotics and Automation*, 2016.
- [19] David Whitney, Eric Rosen, James MacGlashan, Lawson LS Wong, and Stefanie Tellex. Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1006–1013. IEEE, 2017.
- [20] Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *Proceedings of the 9th Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pages 5450–5463. PMLR, 2025. URL <https://proceedings.mlr.press/v305/xiong25a.html>. arXiv:2506.15666.
- [21] Bo Yang, Jian Huang, Xiaolong Li, Xinxing Chen, Caihua Xiong, and Yasuhisa Hasegawa. Natural grasp intention recognition based on gaze fixation in human-robot interaction, 2020. URL <https://arxiv.org/abs/2012.08703>.