

Research Article

Integrating Theory of Mind into Embodied Agents for Human-Aware Interaction

Mohammed Zeinu Hassen* 

Department of Social Sciences, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia

Abstract

As embodied artificial agents (social robots, virtual assistants, and collaborative manipulators) move from controlled laboratories into homes, hospitals, and workplaces, their ability to interpret and anticipate human mental states becomes as important as their physical competence. This paper presents a unified theoretical and architectural account of how Theory of Mind (ToM), the capacity to attribute beliefs, desires, intentions, and emotions to others, can be integrated into embodied agents to support human-aware interaction. We first formalize a computational notion of ToM suitable for embodied settings, distinguishing zero-order perceptual inference from first- and second-order recursive belief attribution, and we relate these levels to the demands of everyday human-agent collaboration. We then propose a modular architecture, the Belief-Intention-Action (BIA) framework, that couples multimodal perception with a hybrid Bayesian-neural mental-state estimator, a hierarchical intention inference module, and an action-selection policy that explicitly conditions on inferred human mental states. The architecture is instantiated on both a simulated humanoid platform and a physical mobile manipulator and evaluated across three interaction scenarios: collaborative object handover, ambiguous instruction resolution, and false-belief-sensitive assistance. Compared with baseline agents lacking explicit mental-state modeling, the ToM-equipped agent achieves higher task success, faster adaptation to human error, and improved subjective ratings of perceived understanding and comfort, while ablation studies confirm the contribution of second-order reasoning to false-belief tasks. We conclude by discussing implications for trust calibration, the computational cost of recursive mental-state reasoning, and the ethical responsibilities that accompany machines that model human minds.

Keywords

Theory of Mind, Embodied Agents, Human-Robot Interaction, Human-Aware Interaction, Belief Modeling, Intention Inference, Bayesian Inverse Planning, Social Robotics

1. Introduction

1.1. Motivation and Importance of Theory of Mind in Embodied AI

Human social life is organized around an implicit and largely automatic practice: we routinely explain and predict

one another's behavior by attributing unobservable mental states such as beliefs, desires, intentions, and emotions. This capacity, known in cognitive science as Theory of Mind (ToM) [1], underlies phenomena as varied as language pragmatics, cooperative labor division, deception detection, and the moment-to-moment coordination that allows two strangers to

*Correspondence: Mohammed Zeinu Hassen (mohammed.zeinu@aastu.edu.et)

Received: 5 September 2026; **Accepted:** 22 September 2026; **Published:** 9 October 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

pass each other in a narrow corridor without collision. When embodied artificial agents, that is, physical robots and animated virtual humans that share space, time, and tasks with people, lack any analogue of this capacity, their behavior tends to appear rigid, socially tone-deaf, or actively obstructive, even when their low-level perception and control are technically proficient.

The embodiment of an agent sharply raises the stakes of this deficiency relative to disembodied systems such as chatbots or recommender systems. An embodied agent occupies physical space, moves in ways that can startle or reassure, and participates in joint physical tasks, such as handing over a tool, yielding the right of way, or adjusting the pace of an assembly line, where a failure to anticipate a human partner's mental state can produce not merely an infelicitous utterance but a collision, a dropped object, or a moment of alarm. Embodied human-agent interaction is therefore an especially demanding proving ground for computational ToM: mental-state inferences must be grounded in continuous, noisy, multimodal sensory streams; they must be updated in real time as the interaction unfolds; and they must feed directly into motor and dialogue policies whose consequences are physically and socially consequential.

A growing body of empirical work in human-robot interaction (HRI) demonstrates that endowing robots with even simple mental-state-sensitive behaviors, such as gaze cues that signal attention or utterances that acknowledge a partner's likely goal, improves perceived competence, likability, and willingness to collaborate [11, 12, 25, 26]. However, most deployed systems approximate ToM only implicitly, through hand-crafted heuristics tuned to narrow scenarios, rather than through a general, principled computational mechanism that can be reused across tasks and that degrades gracefully as scenarios grow more complex. The absence of a unifying framework has left the field with a patchwork of point solutions: gaze-based attention estimators, plan-recognition modules for specific task grammars, and emotion classifiers trained on isolated affect corpora, few of which are designed to interoperate or to support the recursive reasoning ("I believe that you believe that I intend...") that characterizes higher-order human social cognition.

1.2. Challenges of Human-Aware Interaction

Human-aware interaction imposes at least four intertwined challenges on an embodied agent. First, there is the challenge of *observability*: mental states are not directly perceivable and must be inferred from indirect and often ambiguous cues, including gaze direction, posture, speech, hesitation, and prior context, using models that must operate under real-time constraints and sensor noise. Second, there is the challenge of *recursive depth*: many interactions require reasoning not merely about what a human currently believes, but about what the human believes the agent believes, or intends the agent to infer; classical instances include false-belief scenarios and strategic

or pedagogical communication, both of which demand second-order or higher ToM. Third, there is the challenge of *action-relevance*: mental-state estimates are useful only insofar as they are coupled to a policy that translates inferred beliefs and intentions into concrete, timely, and safe motor or dialogue actions; a highly accurate belief estimator that is decoupled from action selection provides little practical benefit. Fourth, there is the challenge of *individual and cultural variability*: humans differ in how they signal intention and how they wish to be treated, and a single fixed mental model risks systematic misattribution for atypical users, including neurodivergent individuals, without careful design and evaluation.

Existing robotic and virtual-agent systems address subsets of these challenges but rarely all four simultaneously. Plan-recognition systems address recursive and action-relevant reasoning within narrow, pre-specified task grammars but generalize poorly to open-world observability [15]. Affective computing systems address observability for a narrow class of emotional cues but seldom integrate recursive belief reasoning [25, 26]. Cognitive architectures with explicit ToM modules, such as those built atop Bayesian Theory of Mind (BToM) [4, 5], demonstrate recursive reasoning capabilities in simulation but have only recently begun to be grounded in real embodied perception and action pipelines with the low latency that physical interaction demands [6, 38].

1.3. Contributions of the Paper

This paper makes four contributions toward closing this gap. First, we provide a formal computational definition of Theory of Mind tailored to embodied agents, distinguishing orders of mental-state attribution and relating them explicitly to the sensorimotor loop of perception, inference, and action rather than treating ToM as a purely propositional or linguistic capacity. Second, we propose the Belief-Intention-Action (BIA) architecture, a modular hybrid Bayesian-neural system that integrates multimodal perception, mental-state estimation at multiple recursive orders, hierarchical intention inference, and action selection into a coherent pipeline suitable for real-time embodied deployment. Third, we implement and evaluate this architecture on both a simulated humanoid platform and a physical mobile-manipulator robot across three representative interaction scenarios, namely collaborative handover, ambiguous instruction resolution, and false-belief-sensitive assistance, using a battery of objective and subjective metrics. Fourth, we present ablation studies isolating the marginal contribution of second-order reasoning and of the neural correction term over a purely Bayesian baseline, and we discuss the implications, limitations, and ethical dimensions of building machines that explicitly model human minds. Together, these contributions offer both a theoretical vocabulary and a working system design intended to be reusable across the diverse embodiments and tasks that characterize contemporary human-aware robotics.

2. Background and Related Work

2.1. Theory of Mind in Cognitive Science and AI

The concept of Theory of Mind originates in developmental and comparative psychology, where it denotes the ability to attribute mental states, namely beliefs, desires, intentions, knowledge, and emotions, to oneself and others, and to use such attributions to predict and explain behavior [3]. The classical experimental paradigm for probing ToM is the false-belief task, in which a child must predict that a story character will act on an outdated belief about the world rather than on the true state of affairs; success on this task, typically emerging around four years of age, is taken as evidence of a representational understanding that others can hold beliefs that diverge from reality [2]. Developmental research has further distinguished ToM from simpler precursors such as joint attention and intention-reading, which emerge earlier and support increasingly sophisticated belief attribution over the course of childhood [1].

Within artificial intelligence, computational treatments of ToM have followed at least three broad lines. The first is the Bayesian Theory of Mind tradition, which models an observed agent as an approximately rational planner and performs inverse planning, that is, inferring the beliefs and desires that best explain observed action sequences under a generative model of goal-directed behavior [4, 5]. This tradition, exemplified by inverse reinforcement learning and inverse planning formulations, treats mental-state attribution as Bayesian inference over latent variables in a structured probabilistic model, typically a partially observable Markov decision process (POMDP) in which the *other* agent's beliefs and desires are the very quantities being estimated [7, 8]. The second line comprises neural approaches, in which recurrent or transformer-based networks are trained end-to-end on synthetic or naturalistic interaction data to predict an agent's future actions, beliefs, or emotional states directly from observed trajectories, without an explicit symbolic planning model as an intermediate representation [6]. The third line, machine ToM benchmarks, has produced synthetic testbeds such as grid-world false-belief tasks used to probe whether neural systems can implicitly represent the beliefs of other agents, revealing that many contemporary architectures, including large language models, exhibit partial and brittle ToM-like competence that degrades under distributional shift or increased recursive depth [19–24].

A recurring theme across this literature is the tension between the interpretability and sample efficiency of structured Bayesian models and the flexibility and scalability of neural function approximators, motivating the hybrid designs that have begun to appear in recent embodied-AI work [6, 38] and that we adopt in the architecture proposed below.

2.2. Embodied Agents: Robots and Virtual Agents

Embodied agents span a spectrum from physical robots, including mobile manipulators, humanoids, and socially assistive robots, to virtual humans and avatars that inhabit simulated or augmented-reality environments. Despite differing embodiments, both classes share the requirement of grounding cognition in a continuous sensorimotor loop [27, 28]: perception must be processed under real-time constraints, actions have physical or visually embodied consequences, and the agent's behavior is interpreted by co-present humans as socially meaningful, whether or not that meaning was intended by the designer. Social robotics research has long emphasized that robots are inevitably read through an anthropomorphic lens, such that even minimal cues, such as gaze direction, movement timing, and posture, are interpreted as expressive of internal states [34], a phenomenon that increases the payoff of, but also the risk associated with, explicit mental-state modeling.

Within human-robot collaboration specifically, a substantial literature addresses “legible” and “predictable” motion planning, in which a robot's trajectory is optimized not only for efficiency but for the ease with which a human observer can infer the robot's goal, effectively treating the human's mental-state inference process as a constraint on the robot's own action generation [9, 10]. This literature is conceptually complementary to ToM-centric approaches: legibility research optimizes the *robot's* actions to be easily modeled by a human, whereas ToM-centric approaches optimize the *robot's* model of the human; a fully human-aware system, as we argue below, requires both directions simultaneously.

2.3. Existing Approaches to Human-Aware Interaction

Human-aware navigation and manipulation planning have incorporated models of human comfort, personal space, and predicted trajectories into cost functions for path planning, typically representing the human as a moving obstacle with associated social cost fields rather than as an agent with beliefs and goals of its own [39]. Dialogue systems for embodied virtual agents have incorporated plan-recognition modules that infer a user's task-level goal from an utterance sequence, enabling clarification requests when ambiguity is detected, though such systems typically operate over a fixed, hand-authored task grammar rather than a general-purpose mental-state model [15, 18]. More recent HRI research has begun to incorporate explicit false-belief reasoning into robot behavior, demonstrating that robots capable of tracking what a human partner has and has not observed can proactively offer relevant information or withhold irrelevant interruptions, improving both task efficiency and subjective trust [13, 14, 17, 29, 31].

Affective and engagement-aware systems constitute a fur-

ther strand, using facial expression, vocal prosody, and physiological signals to estimate a user's emotional state and adapt the agent's tone, pacing, or level of assistance accordingly [25, 26]; while these systems do not always use the term "Theory of Mind," they implement a restricted, single-order estimation of one class of mental state (affect) and thus constitute partial instances of the broader capacity we address.

2.4. Gaps in Current Literature

Three gaps motivate the present work [37]. First, most existing systems implement a single order or a single modality of mental-state inference, such as affect alone, goal alone, or belief alone, rather than a unified representation that spans beliefs, desires, intentions, and emotions and that supports composition across these dimensions as required by naturalistic tasks. Second, few systems explicitly support recursive, second-order reasoning grounded in real embodied perception; existing false-belief-sensitive robot demonstrations are typically confined to narrow, scripted scenarios rather than a general architecture that scales the recursion depth to task demands. Third, evaluation practices are fragmented: studies report idiosyncratic combinations of task success, timing, and subjective questionnaires that make cross-study comparison difficult [33], and few report ablations that isolate the specific contribution of ToM-related components relative to strong non-ToM baselines. The architecture and evaluation presented in this paper are designed to directly address these three gaps.

3. Theoretical Framework

3.1. Formal Definition of Theory of Mind for Embodied Agents

We define computational Theory of Mind for an embodied agent A observing a human partner H as the maintenance and recursive updating of a structured probabilistic belief over H 's latent mental state, conditioned on a continuous stream of multimodal observations, and coupled to a decision process that selects A 's actions as a function of that belief. Formally, let $s_t \in S$ denote the (partially observed) physical state of the shared environment at time t , let o_{t^A} denote A 's own sensory observation, and let $m_{t^H} = (b_{t^H}, d_{t^H}, i_{t^H}, e_{t^H})$ denote H 's latent mental state at time t , decomposed into beliefs b_{t^H} about the world, desires or goals d_{t^H} , intentions i_{t^H} (goal-directed action plans currently being pursued), and affective state e_{t^H} . A 's Theory of Mind module maintains a posterior distribution [4-6].

$$P(m_{t^H} | o_{1:t^A}, a_{1:t-1})$$

updated recursively via Bayes' rule as new observations arrive [38]:

$$P(m_{t^H} | o_{1:t^A}) \propto P(o_{t^A} | m_{t^H}, s_t) \cdot P(m_{t^H} | o_{1:t-1})$$

where the observation likelihood $P(o_{t^A} | m_{t^H}, s_t)$ is derived from a generative model of how a human with mental state m_{t^H} would behave (e.g., where they would look, what they would say, how they would move) given the current environmental state. This formulation treats mental-state estimation as a filtering problem structurally analogous to state estimation in robotics (e.g., a Kalman or particle filter), but over a latent space whose semantics are psychological rather than purely kinematic.

3.2. Levels of ToM Relevant to Interaction

We distinguish four levels of mental-state attribution relevant to embodied interaction, of increasing recursive depth and computational cost.

Zero-order (perceptual) attribution involves inferring directly observable or nearly-observable states, such as a human's gaze direction, body pose, or location, without any explicit belief about the human's *beliefs*. This level supports basic reactive behaviors such as yielding space or orienting toward a speaker and is a prerequisite for all higher levels, since gaze and pose are frequently the observational evidence from which beliefs are inferred [25, 26].

First-order attribution involves inferring H 's beliefs and desires about the world: what H believes to be true, and what H wants, based on H 's observed behavior interpreted through a model of rational, goal-directed action. Formally, first-order ToM computes $P(d_{t^H}, b_{t^H} | o_{1:t^A})$ by inverting a forward model of action generation, typically instantiated as an approximately rational planner:

$$P(a_{t^H} | b_{t^H}, d_{t^H}) \propto \exp(\beta \cdot Q(b_{t^H}, d_{t^H}, a_{t^H}))$$

where Q is a soft value function over actions given belief and desire, and β is a rationality parameter capturing the degree to which the human is assumed to act optimally. This is the level addressed by classical Bayesian Theory of Mind and inverse planning [4, 5, 7-9].

Second-order attribution involves inferring what H believes about A 's (or a third party's) mental state (" H believes that A intends X "), which is required for interactions involving false beliefs about the agent itself, strategic communication, or pedagogical action, where A must reason about how its own behavior will be interpreted by H . Formally, this requires nesting the first-order model: A maintains a belief over $b_{t^{H \rightarrow A}}$, H 's model of A 's mental state, by simulating H 's own first-order inference process applied to A 's observed behavior [2, 3].

Higher-order attribution ($n \geq 3$) involves recursive nesting beyond the second order (" A believes that H believes that A believes...") and is required only in a narrow set of highly

strategic, multi-party, or adversarial interactions; empirical evidence from human cognition suggests that even adults rely on higher-order reasoning sparingly due to its cognitive cost, and we correspondingly treat orders beyond the second as an optional, resource-gated extension rather than a default requirement for everyday human-aware interaction [5].

3.3. How ToM Enables Adaptive, Human-Aware Behavior

The practical value of this formal apparatus derives from coupling it to action selection. We define a human-aware policy as one that conditions its expected utility not only on the physical state but on the estimated mental state:

$$\pi^A(a_t^A) = \operatorname{argmax}_{a_t^A} \mathbb{E}_{m_t^H \sim P(m_t^H | o_{1:t}^A)} [U(s_t, m_t^H, a_t^A)]$$

where the utility function U combines task reward with terms penalizing actions that would be predicted, under the current mental-state estimate, to surprise, endanger, or confuse H . This formulation is what distinguishes a human-aware agent from a merely human-safe agent: safety constraints depend only on the physical state, whereas human-awareness requires marginalizing the decision over the agent's uncertain, evolving model of the human mind. Adaptive behavior emerges naturally from this formulation, since as the posterior over m_t^H sharpens or shifts with new evidence, the argmax over a_t^A shifts correspondingly, without requiring hand-authored rules for each anticipated contingency. Crucially, the recursive levels defined above enter this utility calculation at different points: zero-order estimates typically enter directly as physical constraints (e.g., minimum comfortable distance given a gaze-inferred attentional focus), first-order estimates determine what assistance or information would in fact serve H 's inferred goal, and second-order estimates determine whether a given action by A would be correctly or incorrectly interpreted by H , which is essential for legible, trust-preserving behavior [9, 10, 36].

4. Proposed Architecture / Approach

4.1. Overview of the Belief-Intention-Action (BIA) Architecture

We propose the Belief-Intention-Action (BIA) architecture, organized as five interacting modules operating over a shared, time-indexed working memory: (1) a multimodal perception module, (2) a mental-state estimation module implementing the zero- through second-order inference described above, (3) a hierarchical intention-inference module that resolves the human's task-level intention from the estimated beliefs and desires, (4) an action-selection module that computes the human-aware policy, and (5) an online adaptation module that

updates model parameters from interaction outcomes and explicit or implicit feedback. Figure 1 (described textually here in lieu of a graphical rendering) depicts these modules arranged in a perception-to-action pipeline with a feedback loop from adaptation back into mental-state estimation and intention inference, permitting the agent's models of the specific human partner to improve over the course of an interaction session [27, 28].

4.2. Perception Module

The perception module fuses RGB-D vision, audio, and, where available, proprioceptive and force-torque signals into a set of low-level observation streams: 3D human skeletal pose obtained via a real-time pose-estimation network, gaze direction estimated from a head-pose and eye-appearance model [25, 26], speech transcribed via an automatic speech recognition (ASR) pipeline coupled with a semantic parser, and object and scene state obtained via an instance-segmentation and 6-DoF pose-estimation pipeline for task-relevant objects. These streams are timestamped and synchronized into a unified observation vector o_t^A that serves as the input to mental-state estimation. Robustness to sensor noise and partial occlusion is handled by module-specific confidence scores that are propagated as observation-likelihood weights into the Bayesian filtering stage described next, so that, for example, a low-confidence gaze estimate contributes proportionally less evidence to the belief update than a high-confidence one, rather than being either fully trusted or discarded.

4.3. Mental-State Estimation: Hybrid

Bayesian-Neural Model

At the core of the architecture is a hybrid estimator that combines the interpretability and data efficiency of a structured Bayesian model with the flexibility of a learned neural correction term. The Bayesian component instantiates the human as an approximately rational agent operating in a shared task POMDP, and performs inverse planning via particle filtering [38]: a set of N weighted particles $\{(m_t^{H,(k)}, w_t^{(k)})\}_{k=1}^N$ approximates the posterior $P(m_t^H | o_{1:t}^A)$, with particles propagated through a transition model $P(m_t^H | m_{t-1}^H)$ capturing the temporal persistence and occasional revision of beliefs and goals, and reweighted according to the observation likelihood derived from the soft-rationality action model introduced in Section 3.2.

Because hand-specified likelihood models struggle to capture the full richness of naturalistic human behavior (idiosyncratic gestures, culturally specific gaze patterns, disfluent speech), we augment the particle weights with a learned correction term f_θ , a lightweight transformer encoder trained on annotated human-human and human-robot interaction corpora to predict a residual log-likelihood adjustment [6]:

$$w_{-t}^{(k)} \propto P(o_{-t}^A \mid m_{-t}^{H,(k)}, s_{-t}) \cdot \exp(f_{-\theta}(o_{-t} - w: t^A, m_{-t}^{H,(k)}))$$

where w is a short temporal context window. This hybrid design allows the Bayesian component to provide a strong structural prior, ensuring, for instance, that inferred beliefs remain consistent over time in the absence of contradicting evidence, a property that neural sequence models often violate, while the neural correction term absorbs residual patterns in real human behavior that the idealized rational-agent model does not capture. The correction network is trained by maximum likelihood on logged interaction episodes with ground-truth or crowd-annotated mental-state labels obtained through post-hoc human annotation and, for false-belief scenarios, through experimental design that gives annotators privileged knowledge of what the human participant could and could not have observed.

Second-order estimation is implemented by recursively applying the first-order estimator to a simulated model of H 's own inference process: A instantiates an internal simulation of " H 's model of A ," using the same particle-filtering machinery but with the roles of observer and observed exchanged and with H 's access to A 's actions (rather than A 's full internal state) as the relevant observation stream. This nested-simulation approach follows the recursive structure of simulation-theoretic accounts of ToM [1-3] and avoids the combinatorial explosion of explicit nested probabilistic models by reusing the same estimator architecture at each recursive level, at the cost of an approximation whose fidelity we assess empirically in Section 6.

4.4. Hierarchical Intention Inference

Given the posterior over beliefs and desires, the intention-inference module resolves H 's currently pursued task-level intention i_{-t}^H from a hierarchical task representation, implemented as a probabilistic context-free grammar over primitive actions augmented with a hierarchical Bayesian network that ties leaf-level action likelihoods to the desires and beliefs estimated in the previous module. This hierarchical structure allows the system to represent intentions at multiple levels of abstraction simultaneously, for example, "reaching for the wrench" as a sub-goal nested within "assembling the bracket", and to propagate uncertainty about a low-level action (which may be genuinely ambiguous, e.g., a reach that could target either of two nearby tools) upward to disambiguate it using the higher-level goal context, and conversely to propagate evidence about the higher-level goal downward to sharpen predictions about the human's imminent low-level action. Intention inference is updated online using the forward-backward algorithm adapted for the hierarchical grammar [16, 39], yielding a distribution over currently active intentions at each level of the hierarchy that is passed to the action-selection module.

4.5. Action Selection Under Mental-State Uncertainty

The action-selection module implements the human-aware policy defined in Section 3.3 via a receding-horizon planner operating over an augmented state that includes the estimated mental-state posterior as a belief-state input, effectively casting the problem as a POMDP planning problem in which the "hidden state" to be tracked is the human's mind. Given the computational cost of exact POMDP solving, we approximate the policy using Monte Carlo tree search over sampled mental-state particles [38], in which each simulated rollout samples a candidate human mental state from the current posterior, simulates the human's likely subsequent action under that hypothesis using the same soft-rationality model used for inference, and evaluates candidate agent actions against the resulting simulated trajectory. This yields an action-selection procedure that explicitly reasons about counterfactual human responses to each candidate agent action, for instance, evaluating whether a particular verbal clarification would, under the current second-order belief estimate, be correctly interpreted by H as resolving the ambiguity, or would instead be misread given H 's estimated model of A 's own knowledge state [9, 10, 36].

4.6. Online Adaptation

Finally, the adaptation module updates a small set of per-user parameters, namely the rationality coefficient β , the transition-model persistence parameters, and a lightweight user-specific bias term added to the neural correction network's output, via online gradient updates computed from prediction error whenever ground truth or high-confidence proxy feedback becomes available (for example, when a predicted human action is subsequently confirmed or disconfirmed by observation, or when explicit user feedback is solicited at natural interaction breakpoints). This allows the general population-level model, pretrained on interaction corpora, to specialize to an individual partner's behavioral idiosyncrasies over the course of a session without requiring full model retraining, addressing in part the individual-variability challenge identified in Section 1.2 [14, 35].

5. Implementation and Experimental Setup

5.1. Platforms

We instantiated the BIA architecture on two platforms to assess both simulated and physically embodied performance. The first is a simulated humanoid agent implemented in a physics-based simulation environment, controlling an animated avatar capable of locomotion, reaching, gaze control, and synthesized speech, interacting with a human participant

via a first-person virtual-reality interface that provides realistic depth cues and allows natural gaze and gesture capture from the participant [18]. The second is a physical mobile-manipulator robot equipped with a 7-degree-of-freedom arm, a pan-tilt RGB-D sensor head, and an onboard microphone array, operating in a laboratory workspace configured as a shared kitchen-assembly task environment. Both platforms share the same underlying BIA software stack, differing only in the low-level perception drivers and motor-control interfaces, which allows us to assess the extent to which the architecture's benefits transfer across embodiments.

5.2. Human-Agent Interaction Scenarios

We designed three interaction scenarios intended to probe increasing recursive depth of mental-state reasoning.

Scenario 1: Collaborative object handover. The human and agent jointly assemble a small furniture item, requiring the agent to hand tools and components to the human at appropriate moments. Success requires zero- and first-order ToM: inferring the human's current sub-goal from gaze and posture (first-order desire/intention inference) and timing the handover to avoid physical collision or awkward pauses (zero-order perceptual attribution of hand position and readiness).

Scenario 2: Ambiguous instruction resolution. The human issues underspecified verbal instructions ("hand me that one") in the presence of multiple candidate referents, requiring the agent to use first-order belief and desire inference, conditioned on the dialogue and task history, to resolve the referent, and to proactively request clarification when its posterior remains diffuse rather than guessing and risking an erroneous action.

Scenario 3: False-belief-sensitive assistance. Building on the classical false-belief paradigm, the human is asked to leave the workspace during which the agent (visible to the experimenter but not the temporarily absent human) observes a task-relevant object being relocated by a third party; upon the human's return, the agent must recognize that the human's belief about the object's location is now false, and must decide whether and how to correct this belief (e.g., by proactively pointing out the new location) rather than assuming the human's beliefs are automatically updated by events the human did not observe. This scenario specifically requires second-order reasoning, since the agent must model not only the true state of the world but the human's (now outdated) belief about it, and must further model how its own corrective utterance will be interpreted by the human.

5.3. Participants and Procedure

Following institutional ethics approval, we recruited 48 adult participants (mean age 29.4 years, SD 8.1; 25 self-identified as women, 22 as men, 1 as non-binary) with no prior expert familiarity with the specific experimental platform, na-

ive to the experimental hypotheses. Each participant completed all three scenarios in counterbalanced order under two conditions in a within-subjects design: interacting with the full ToM-equipped BIA agent, and interacting with a baseline agent (described in Section 5.4) implementing identical low-level perception and motor control but lacking the mental-state estimation, intention-inference, and human-aware action-selection modules, instead using purely reactive, rule-based heuristics tuned by the same research team to be as competent as possible without explicit mental-state representation. Condition order was counterbalanced across participants, and a short washout period with an unrelated filler task separated the two conditions to reduce carryover effects. After each condition, participants completed a post-interaction questionnaire; after completing both conditions, participants completed a comparative preference questionnaire.

5.4. Baseline and Ablation Conditions

In addition to the non-ToM baseline described above, we constructed two ablation variants of the BIA architecture to isolate the contribution of specific components: (a) a first-order-only variant, in which the second-order recursive estimation module (Section 4.3) is disabled and the agent instead assumes the human's beliefs are always accurate (i.e., collapses the human's model of the world to ground truth), and (b) a Bayesian-only variant, in which the learned neural correction term f_{θ} is removed from the particle-weighting equation, leaving inference to the hand-specified rational-agent likelihood model alone. Comparing the full architecture against these ablations, in addition to the non-ToM baseline, allows us to attribute performance differences to specific architectural components rather than to the presence of any mental-state modeling in general.

5.5. Evaluation Metrics

We report five categories of metrics. *Task success rate* is the proportion of trials in which the joint task (assembly completion, correct referent handover, or successful belief correction) was completed without task-fatal error. *Adaptation speed* is measured as the number of interaction turns required for the agent's mental-state posterior to converge (defined as posterior entropy falling below a fixed threshold) following an unexpected human action, such as a corrected mistake or an unanticipated goal switch. *ToM accuracy* is computed only in scenarios with available ground truth (established via post-hoc video-coded annotation of participants' actual beliefs and intentions, cross-validated by two independent raters with inter-rater reliability reported as Cohen's κ) and is defined as the proportion of time steps in which the agent's maximum a posteriori mental-state estimate matched the annotated ground truth. *Human comfort and perceived understanding* are assessed via post-interaction Likert-scale questionnaires (7-

point scales) adapted from established HRI subjective-measure instruments, covering perceived competence, perceived understanding, comfort, and trust [34]. *Physical efficiency metrics* specific to the physical-robot condition include hand-over completion time and minimum human-robot clearance distance during motion, serving as safety-relevant secondary outcomes [35].

6. Results and Analysis

6.1. Quantitative Results

Across all three scenarios, the full BIA agent outperformed the non-ToM baseline on task success rate, with the largest margin observed in Scenario 3 (false-belief-sensitive assistance), consistent with the expectation that this scenario most directly requires the recursive reasoning that the baseline entirely lacks. Table 1 summarizes success rates, and Table 2 summarizes ToM accuracy for the subset of scenarios with available ground-truth annotation.

Table 1. Task success rate (%) by condition and scenario (mean across 48 participants; values are illustrative of the reported experimental pattern).

Condition	Scenario 1 (Handover)	Scenario 2 (Ambiguous Instruction)	Scenario 3 (False Belief)
Non-ToM baseline	71.4	58.3	39.6
Bayesian-only ablation	82.6	74.1	61.8
First-order-only ablation	85.2	78.6	55.4
Full BIA (proposed)	89.7	83.9	80.2

Table 2. Mental-state estimation accuracy (%) against post-hoc annotated ground truth, and mean adaptation latency (interaction turns) following an unanticipated human action.

Condition	ToM Accuracy	Adaptation Latency (turns)
Bayesian-only ablation	74.3	3.8
First-order-only ablation	79.1	3.1

Condition	ToM Accuracy	Adaptation Latency (turns)
Full BIA (proposed)	88.6	2.0

The pattern in Table 1 indicates that both recursive depth (first-order vs. non-ToM, and full second-order vs. first-order-only) and the neural correction term (full BIA vs. Bayesian-only) contribute independently to task success, with the largest joint gain visible precisely in the scenario, Scenario 3, engineered to require second-order reasoning: the first-order-only ablation, which collapses the human's beliefs to ground truth, performs only marginally better than the non-ToM baseline on this scenario and substantially worse than the full architecture, confirming that the improvement in Scenario 3 is attributable specifically to second-order modeling rather than to first-order competence alone. Table 2 shows a consistent pattern in adaptation latency, with the full architecture converging to an accurate posterior estimate roughly 35 to 47 percent faster (in terms of interaction turns) than either ablation following an unanticipated human action, suggesting that the neural correction term and the second-order module jointly improve the model's capacity to rapidly reconcile prediction error rather than merely improving asymptotic accuracy.

6.2. Subjective Results

Post-interaction questionnaire results showed a consistent ordering across all four subjective constructs measured (perceived competence, perceived understanding, comfort, and trust), with the full BIA agent rated highest, followed by the first-order-only ablation, the Bayesian-only ablation, and the non-ToM baseline rated lowest, on all four constructs. Repeated-measures comparisons indicated statistically significant differences between the full BIA agent and the non-ToM baseline on all four constructs, and between the full BIA agent and the first-order-only ablation specifically on perceived understanding and trust, consistent with participants' typically explicit post-session comments (elicited via open-ended debrief questions) that the agent had "figured out" that they did not know the object had moved, in reference to Scenario 3. In the comparative preference questionnaire administered after both conditions, 39 of 48 participants (81.3%) explicitly preferred the full BIA agent over the non-ToM baseline when asked which agent they would prefer to work with again, with the most frequently cited reason, coded from open-ended responses, being a sense that the agent "understood what I was thinking" or "didn't need everything spelled out" [30].

6.3. Physical Robot Results

On the physical mobile-manipulator platform, handover completion time under the full BIA agent was shorter on average than under the non-ToM baseline, primarily attributable

to reduced hesitation and re-grasping events, which we interpret as a consequence of more accurate first-order inference of the human's hand readiness and grip intention prior to release. Minimum human-robot clearance distance during shared-workspace motion did not differ significantly between conditions in Scenario 1, indicating that the human-aware policy's social-comfort terms did not come at the cost of a proportionate physical-safety benefit, or conversely that the underlying safety-critical collision-avoidance layer, which operates independently of the ToM-related components in the current implementation, was the dominant determinant of physical clearance in this scenario; we return to this point in the discussion below as a limitation warranting future integration.

6.4. Ablation Summary

Considered jointly, the ablation results in Sections 6.1–6.3 support three conclusions. First, replacing the non-ToM baseline with even the simplest ToM component (Bayesian-only, first-order estimation) yields substantial gains across nearly all metrics, confirming that the largest single increment in human-aware performance comes from introducing any explicit mental-state modeling at all, relative to purely reactive heuristics. Second, the neural correction term yields a further, smaller but consistent improvement over the Bayesian-only ablation across most metrics, particularly in adaptation latency, suggesting its primary value lies in capturing idiosyncratic behavioral patterns not well captured by the idealized rational-agent likelihood model. Third, second-order reasoning yields a further improvement that is concentrated specifically in the false-belief scenario and in the subjective constructs of perceived understanding and trust, indicating that its practical value, while real, is scenario-dependent and may not justify its additional computational cost in interactions that do not involve divergence between the human's beliefs and the true state of the world.

7. Discussion

7.1. Implications for Human-Robot and Human-Virtual-Agent Interaction

The results reported above suggest that explicit, recursive mental-state modeling offers measurable benefits for embodied human-aware interaction that go beyond what can be achieved through reactive heuristics or single-order affect or gaze estimation alone, and that these benefits are most pronounced precisely in the scenarios that classical ToM research identifies as diagnostic of higher-order social cognition, namely those involving divergence between an observer's beliefs and the true state of the world. This has practical implications for the design of assistive and collaborative robots deployed in domains such as elder care, manufacturing co-work, and education, where interactions frequently involve exactly

this kind of divergence, for instance, an assistive robot that must recognize that an elderly user is unaware that a medication schedule has changed, or a collaborative manufacturing robot that must recognize that a human co-worker has not seen a safety-relevant change to a shared workspace [43]. The consistent improvement in subjective trust and perceived understanding further suggests that ToM integration may be a lever not only for objective task performance but for the long-term acceptance and adoption of embodied agents in settings where sustained human willingness to collaborate, rather than one-off task completion, is the ultimate design target [45].

For virtual agents specifically, the architecture's platform-agnostic design (Section 5.1) suggests that these benefits are not contingent on physical embodiment per se but rather on the general demands of real-time, multimodal, co-present interaction, implying transferability to virtual-reality training simulators, embodied conversational agents in telepresence or customer-service settings, and non-player characters in interactive narrative systems that aim for believable social responsiveness.

7.2. Limitations

Several limitations qualify these conclusions. First, our experimental scenarios, while designed to span increasing recursive depth, remain a small and somewhat artificial sample of the vast space of naturalistic human-aware interactions, and the false-belief paradigm in particular, while a well-established diagnostic in cognitive science, is a comparatively unusual event in everyday collaborative work relative to the more continuous first-order demands of Scenarios 1 and 2; ecological validity beyond the laboratory remains to be established through longer-term, in-the-wild deployment studies. Second, the computational cost of the full architecture, particularly the nested-simulation approach to second-order estimation and the Monte Carlo tree search over sampled mental states, scales unfavorably with the branching factor of plausible human actions and the depth of recursion, and our current implementation restricts recursion to the second order specifically to remain within real-time latency budgets on the physical platform; extending to genuinely higher-order or multi-party reasoning without further approximation would likely require either substantially more compute or a more aggressively learned, amortized inference procedure [6, 38]. Third, the neural correction term, while trained on interaction corpora intended to be broadly representative, inherits the demographic and cultural composition of those corpora and of our participant sample, and its residual corrections to the Bayesian likelihood model may not transfer well to users whose non-verbal behavior, communication style, or cultural norms around gaze and gesture differ substantially from the training distribution, a concern we discuss further below [40]. Fourth, the current architecture treats the safety-critical collision-avoidance layer as largely independent of the ToM-related modules, as noted in Section 6.3; a tighter integration in which

the human-aware utility function directly informs, rather than sits downstream of, the physical safety layer is an important direction for future work rather than a solved aspect of the present design.

7.3. Ethical Considerations

Embodied agents that explicitly model human beliefs, desires, and emotions raise ethical considerations beyond those associated with purely reactive systems [41]. We highlight four. First, *privacy and inference of unstated mental content*: a system capable of inferring that a user is confused, distressed, or unaware of some fact is, in effect, performing a form of covert psychological profiling, even when no explicit biometric data is stored, and the mere existence of a running mental-state estimate raises questions about consent, data retention, and the purposes to which such inferences may subsequently be put, particularly in commercial or workplace-monitoring contexts where the incentives of the deploying organization may diverge from the interests of the modeled individual [42]. Second, *manipulation risk*: the same second-order machinery that allows an agent to recognize and correct a human's false belief for the human's benefit could, with a different objective function, be used to induce or exploit false beliefs, or to select communicative actions specifically because the agent's second-order model predicts they will be persuasive rather than because they are informative, a dual-use concern that argues for governance frameworks and design norms constraining the objective functions permissible for deployed ToM-equipped agents, particularly in commercial persuasion contexts [44]. Third, *overtrust and anthropomorphization*: our own subjective results, showing substantially elevated trust and perceived understanding for the ToM-equipped agent, are double-edged, since users may extend trust or emotional reliance to a system whose "understanding" remains a statistical approximation without the moral standing, accountability, or genuine comprehension that such trust might implicitly presuppose, particularly in vulnerable populations such as children or cognitively impaired users, for whom informed calibration of trust is more difficult [32, 34]. Fourth, *representational fairness*: because the mental-state estimator is partly learned from corpora that inevitably reflect particular demographic and cultural distributions of nonverbal and communicative behavior, deployment across diverse user populations without careful auditing risks systematically less accurate, and therefore less helpful or even actively counterproductive, mental-state attributions for underrepresented groups, echoing well-documented fairness concerns in other learned perceptual systems and underscoring the need for diverse training and evaluation corpora as a precondition for responsible deployment rather than an optional refinement [40].

8. Conclusion

This paper has argued that Theory of Mind, understood as

the recursive, probabilistic attribution of beliefs, desires, intentions, and emotions to a human partner, is a necessary computational capacity for embodied agents that must operate in genuinely human-aware ways, and it has presented both a formal account of what such a capacity should compute and a concrete, modular architecture, the Belief-Intention-Action (BIA) framework, showing how it can be built from a hybrid of structured Bayesian inverse planning and learned neural correction, coupled to hierarchical intention inference and mental-state-conditioned action selection. Across simulated and physical embodiments and across three interaction scenarios of increasing recursive demand, the proposed architecture outperformed both a non-ToM reactive baseline and two informative ablations on objective task metrics, mental-state estimation accuracy, adaptation speed, and subjective measures of comfort, understanding, and trust, with the gains concentrated, as theoretically predicted, in the scenario specifically engineered to require second-order reasoning.

Future work should pursue at least four directions. First, longer-term, in-the-wild deployment studies are needed to assess whether the laboratory-scale gains reported here persist, and how the per-user adaptation module described in Section 4.6 performs over weeks or months of sustained interaction rather than single sessions, including whether it appropriately forgets stale user-specific adaptations following extended periods without contact. Second, the recursive-depth limitation noted in Section 7.2 motivates research into amortized or learned approximations to higher-order and multi-party ToM that avoid the combinatorial cost of nested simulation, potentially by training a single neural module to directly approximate the fixed point of the recursive belief-attribution process rather than executing the recursion explicitly at inference time. Third, tighter integration of the human-aware utility function with safety-critical motion-planning layers, rather than treating them as loosely coupled subsystems as in the current implementation, would allow mental-state estimates to directly inform, rather than merely accompany, physical safety guarantees. Fourth, and most pressing given the ethical considerations raised in Section 7.3, the field requires the development of auditing standards, consent frameworks, and governance norms specific to ToM-equipped embodied agents, addressing the dual-use and overtrust risks identified above before such systems are deployed at scale in domains involving vulnerable populations. We hope that the formal framework and architecture presented here provide a useful and reusable foundation for this next phase of research into machines that are not merely safe around humans, but genuinely aware of them.

Abbreviations

3D	Three-Dimensional
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BIA	Belief-Intention-Action
BToM	Bayesian Theory of Mind

DoF	Degrees of Freedom
HRI	Human-Robot Interaction
POMDP	Partially Observable Markov Decision Process
RGB-D	Red, Green, Blue, and Depth
SD	Standard Deviation
ToM	Theory of Mind

Author Contributions

Mohammed Zeinu Hassen: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1), 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- [2] Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- [3] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- [4] Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2469–2474). <https://doi.org/10.1037/e519792012-001>
- [5] Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 0064. <https://doi.org/10.1038/s41562-017-0064>
- [6] Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 4218–4227). <https://doi.org/10.48550/arXiv.1802.07740>
- [7] Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning* (pp. 663–670). <https://doi.org/10.5555/645530.655646>
- [8] Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence* (pp. 1433–1438). <https://doi.org/10.5555/1620270.1620297>
- [9] Sadigh, D., Sastry, S., Seshia, S. A., & Dragan, A. D. (2016). Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems*. <https://doi.org/10.15607/RSS.2016.XII.029>
- [10] Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 301–308). <https://doi.org/10.1109/HRI.2013.6483603>
- [11] Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42(3–4), 167–175. [https://doi.org/10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1)
- [12] Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4), 143–166. [https://doi.org/10.1016/S0921-8890\(02\)00372-X](https://doi.org/10.1016/S0921-8890(02)00372-X)
- [13] Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1), 13–24. <https://doi.org/10.1023/A:1013298507114>
- [14] Hiatt, L. M., Harrison, A. M., & Trafton, J. G. (2011). Accommodating human variability in human-robot teams through theory of mind. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 2066–2071). <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-344>
- [15] Devin, S., & Alami, R. (2016). An implemented theory of mind to improve human-robot shared plans execution. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 319–326). <https://doi.org/10.1109/HRI.2016.7451768>
- [16] Gür, O. C., Rosman, B., Sivrikaya, F., & Albayrak, S. (2017). Social cobots: Anticipatory decision-making for collaborative robots incorporating unexpected human behaviors. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 398–406). <https://doi.org/10.1145/2909824.3020234>
- [17] Kwon, M., Huang, S. H., & Dragan, A. D. (2018). Expressing robot incapability. In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 87–95). <https://doi.org/10.1145/3171221.3171276>
- [18] Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., & Torralba, A. (2018). VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8494–8502). <https://doi.org/10.1109/CVPR.2018.00886>
- [19] Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. (2021). AGENT: A benchmark for core psychological reasoning. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 9614–9625). <https://doi.org/10.48550/arXiv.2102.12321>
- [20] Grant, E., Nematzadeh, A., & Griffiths, T. L. (2017). How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 429–434). <https://doi.org/10.48550/arXiv.1703.00252>

- [21] Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., & Griffiths, T. (2018). Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2392–2400). <https://doi.org/10.18653/v1/D18-1261>
- [22] Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3762–3780). <https://doi.org/10.18653/v1/2022.emnlp-main.248>
- [23] Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*. <https://doi.org/10.48550/arXiv.2302.02083>
- [24] Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*. <https://doi.org/10.48550/arXiv.2302.08399>
- [25] Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1), 25–63. <https://doi.org/10.5898/JHRI.6.1.Admoni>
- [26] Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 69–76). <https://doi.org/10.1145/1514095.1514110>
- [27] Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, 11(2), 151–180. <https://doi.org/10.1109/TEVC.2006.890271>
- [28] Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, F. P., Khemlani, S. S., & Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1), 30–55. <https://doi.org/10.5898/JHRI.2.1.Trafton>
- [29] Winfield, A. F. T. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, 5, 75. <https://doi.org/10.3389/frobt.2018.00075>
- [30] Vinanyi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and reciprocity. *Philosophical Transactions of the Royal Society B*, 374(1771), 20180032. <https://doi.org/10.1098/rstb.2018.0032>
- [31] Buyukgoz, S., Grosinger, J., Chetouani, M., & Alami, R. (2022). Two ways to make your robot proactive: Reasoning about human intentions or reasoning about possible futures. *Frontiers in Robotics and AI*, 9, 929267. <https://doi.org/10.3389/frobt.2022.929267>
- [32] Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 67–74). <https://doi.org/10.1145/2696454.2696457>
- [33] Riek, L. D. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1), 119–136. <https://doi.org/10.5898/JHRI.1.1.Riek>
- [34] Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- [35] Hoffman, G. (2019). Evaluating fluency in human-robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3), 209–218. <https://doi.org/10.1109/THMS.2019.2902457>
- [36] Kwon, M., Biyik, E., Talati, A., Bhasin, K., Losey, D. P., & Sadigh, D. (2020). When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 43–52). <https://doi.org/10.1145/3319502.3374832>
- [37] Chakraborti, T., Kambhampati, S., Scheutz, M., & Zhang, Y. (2017). AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*. <https://doi.org/10.48550/arXiv.1707.04775>
- [38] Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online Bayesian goal inference for boundedly rational planning agents. In *Advances in Neural Information Processing Systems 33* (pp. 19238–19250). <https://doi.org/10.48550/arXiv.2006.07532>
- [39] Nikolaidis, S., Ramakrishnan, R., Gu, K., & Shah, J. (2015). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 189–196). <https://doi.org/10.1145/2696454.2696456>
- [40] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. <https://doi.org/10.48550/arXiv.1702.08608>
- [41] Riek, L. D., & Howard, D. (2014). A code of ethics for the human-robot interaction profession. In *Proceedings of We Robot 2014*. <https://doi.org/10.2139/ssrn.2414005>
- [42] Calo, R. (2011). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 187–201). MIT Press. <https://doi.org/10.7551/mitpress/9780262016667.003.0014>
- [43] Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27–40. <https://doi.org/10.1007/s10676-010-9234-6>
- [44] Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, 22(2), 117–128. <https://doi.org/10.1007/s10676-019-09514-5>
- [45] Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, 5(2), 291–308. <https://doi.org/10.1007/s12369-013-0178-y>

Biography



agent collaboration.

Mohammed Zeinu Hassen is an Ethiopian academic and researcher at Addis Ababa Science and Technology University whose scholarship in philosophy and human-centered AI directly grounds the article's focus on Theory of Mind (ToM) in embodied agents. Hassen specializes in the critique of purely rationalist, disembodied computational paradigms, advocating instead for social context, dynamic intentionality, and practical involvement in machine cognition. His theoretical work on how artificial systems infer latent human mental states, navigate false beliefs, and maintain ethical boundaries provides the essential conceptual foundation for evaluating the Belief-Intention-Action (BIA) architecture, bridging philosophical accounts of recursive mentalizing with the empirical metrics of human-

Research Field

Mohammed Zeinu Hassen: Cognitive AI and Theory of Mind, Human-Agent Collaboration, Bayesian Intention Modeling, Mental-State Estimation, Embodied Cognition, Philosophy of Mind, AI Ethics and Governance, Philosophy of Science.