PROPOSING A MULTI-AGENT SIMULATION FRAMEWORK FOR QUANTIFYING THE EMERGENCE OF THEORY OF MIND IN FOUNDATION ARTIFICIAL INTELLIGENCE MODELS

Alexandru PÎRJAN[1]
Dana-Mihaela PETROŞANU[2]

Abstract

The ability to reason about others' mental states, known as Theory of Mind (ToM), is an important aspect of social intelligence. Nevertheless, quantifying its emergence in foundation AI models has been hampered by methodological flaws in existing static benchmarks. Current evaluations often rely on single-turn vignettes that are susceptible to dataset contamination and shortcut solutions, making it difficult to distinguish genuine belief reasoning from superficial pattern matching. This paper proposes a multi-agent simulation framework to address this challenge. We embed foundation models as agents in interactive, partially observable environments where success is causally contingent on tracking, inferring, and manipulating others' beliefs. The framework operationalizes ToM through dynamic task families, including false-belief, deception, and coordination games, that make nested belief reasoning very important and necessary for achieving goals. Significantly, this aspect goes beyond purely behavioral outcomes by instrumenting process evidence, such as the alignment of an agent's reported beliefs with ground-truth epistemic states and its subsequent actions. The proposed approach separates intrinsic model competence from the effects of performance artifacts and inference-time scaffolding. The primary contributions are operational semantics for ToM in artificial agents, a suite of belief-critical tasks, and a measurement protocol that integrates outcome success with process-level validation. Applying this framework reveals that while current models exhibit measurable and partially robust first-order ToM-like capabilities, higher-order reasoning is more fragile and strongly dependent on model scale, context, and structured deliberation. By providing an auditable, reproducible, and adversarial robust evaluation substrate, this work establishes a principled methodology for quantifying how and when social cognition emerges in artificial agents, moving the field from unreliable claims toward a cumulative, empirical science.

Keywords: Theory of Mind, Foundation Models, Multi-Agent Simulation, Emergent Abilities, Artificial Social Intelligence, Computational Cognitive Science

JEL Classification: C6, O3, O33, O34, O35, O36

[1] PhD Hab. Full Professor, School of Computer Science for Business Management, Romanian-American University, 1B, Expozitiei Blvd., district 1, code 012101, Bucharest, Romania, alexandru.pirjan@rau.ro, corresponding author

[2] PhD Lecturer, Department of Mathematics-Informatics, National University of Science and Technology Politehnica Bucharest, 313, Splaiul Independentei, district 6, code 060042, Bucharest, Romania, dana.petrosanu@upb.ro

Article's total number of pages: 45

1. Introduction

In order to engage in rich social interactions, artificial agents must be capable of attributing beliefs, desires, and intentions to others. This capacity, known as Theory of Mind (ToM), is an important element of human social cognition [1]. A pressing question is whether large-scale foundation models, trained on broad internet data, exhibit a comparable ability and, if so, under what conditions and through which mechanisms. Early claims of ToM-like skills in such models have been contested, with critiques suggesting that success on evaluation tasks could arise from dataset artifacts, prompting strategies, or superficial pattern matching rather than genuine belief reasoning. Therefore, the field faces a dual challenge, namely, to establish careful operational definitions of ToM suitable for artificial agents and to create evaluation environments that elicit true belief reasoning instead of allowing shortcut solutions. This work addresses both needs by proposing a framework for measuring when and how social-cognitive skills emerge in AI.

1.1 Motivation and research questions

The drive to quantify ToM in foundation models is motivated by theoretical, methodological, and practical considerations [2,3]. Theoretically, it explores whether large-scale, self-supervised training on human data can induce the latent structures necessary for belief reasoning, a subject of open debate in cognitive science. Methodologically, existing ToM tests for AI often contain serious pitfalls, many rely on simple story puzzles or prompted questions that models might solve via shortcuts, casting doubt on claims of genuine mental state attribution. Practically, foundation models are increasingly deployed in interactive roles where they must infer user intent, negotiate with other agents, or coordinate under partial information. Robust measures of social reasoning are therefore essential to ensure these systems behave safely and predictably [4,5]. Guided by these motivations, our research investigates whether foundation models exhibit ToM-like competencies and, if so, under what conditions. We examine how this capability scales with model size or improves with interaction. A very important goal is to design tasks that necessitate belief reasoning, thereby revealing its presence or absence with high fidelity. Ultimately, our framework is engineered to distinguish genuine belief attribution from clever but still shallow heuristics or memorized patterns.

1.2 Contributions and scope

This article offers a unified conceptual and experimental framework for studying the emergence of ToM in foundation models [2,3], comprising construct definition, task creation, and rigorous evaluation . Our first contribution is to operationalize ToM for artificial agents by providing precise definitions for different levels of belief reasoning and by clarifying what constitutes evidence of such capabilities. In order to elicit and test these abilities, we introduce a suite of interactive multi-agent simulation tasks, including false-belief scenarios and deception games, designed so that success requires the inference and manipulation of others' hidden mental states. This design prevents solutions based on superficial pattern matching. In complementing these tasks, we develop a comprehensive measurement protocol that combines outcome-based metrics with process-based analyses of an agent's decision-making, offering deeper insight into its reasoning. Finally, through extensive experiments, we map the conditions under which ToM-like abilities appear in foundation models of different scales. Our empirical findings show that while current models exhibit some robust first-order ToM, higher-order reasoning remains fragile, and we identify the factors that influence this performance. Together, these contributions advance the study of social cognition in machines and lay the groundwork for a more rigorous science of machine ToM.

1.3 Definitions of ToM, emergence, foundation models

In order to ensure clarity, this work adheres to precise definitions for its core concepts. We define ToM as an agent's operational ability to represent and reason about the mental states of others, using these representations to inform its actions. This capacity is treated as a graded property, observable through behavior, rather than as a monolithic trait. The term emergence refers to qualitatively new capabilities that manifest at a certain scale of model complexity, surpassing extrapolations from smaller models. In order for a capability to be considered emergent, it must be reproducible and robust. We also distinguish an agent's intrinsic competence from performance enhanced by external scaffolding, such as memory buffers or planners. Foundation models are the large-scale, pre-trained neural networks that serve as the cognitive core for the agents in our simulations. Our framework explicitly separates the abilities of the base model from the contributions of its surrounding architecture.

1.4 Overview of approach and findings

The subsequent sections of this paper systematically unfold our framework (Figure 1). We begin by situating our work within the existing literature before formally defining the problem and our central hypotheses. We then detail the multi-agent simulation environment and agent architecture developed to test these hypotheses. Following this, we describe the specific evaluated foundation models and the comprehensive protocol used for their assessment, including our proposed metrics and statistical methods. Moving forward, the paper presents our experimental results, makes a deeper analysis of model behaviors and failure modes, and puts forward a discussion of the broader implications, limitations, and ethical considerations of our findings. We conclude by detailing our commitment to reproducibility and outlining future research directions.

Our results provide a nuanced picture of ToM in current foundation models. We find that agents can achieve non-trivial performance on tasks requiring first-order belief reasoning, such as tracking another agent's knowledge to succeed in a cooperative endeavor. Nevertheless, robust higher-order ToM, which involves reasoning about nested beliefs, remains largely unattainable with current capabilities. While larger models and specific scaffolding techniques yield performance gains, we also demonstrate that these apparent abilities are often fragile, diminishing significantly under adversarial or out-of-distribution conditions. These findings emphasize both the promise and the current limitations of artificial social cognition, reinforcing the very important need for the principled evaluation framework we propose [6].

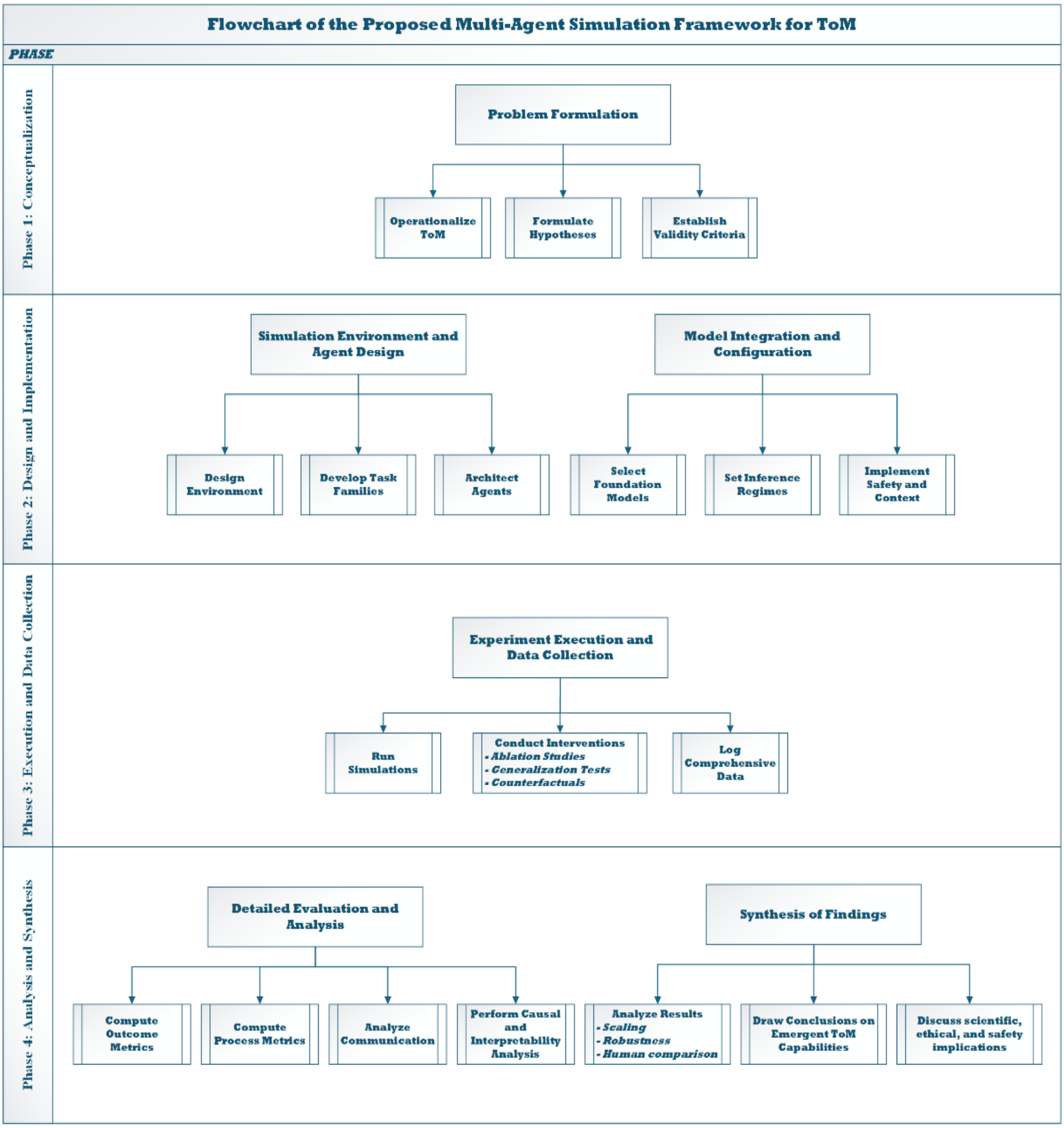
Flowchart of the Proposed Multi-Agent Simulation Framework for ToM
PHASE
Phase 1: Conceptualization
Problem Formulation
Operationalize ToM
Formulate Hypotheses
Establish Validity Criteria
Phase 2: Design and Implementation
Simulation Environment and Agent Design
Design Environment
Develop Task Families
Architect Agents
Model Integration and Configuration
Select Foundation Models
Set Inference Regimes
Implement Safety and Context
Phase 3: Execution and Data Collection
Experiment Execution and Data Collection
Run Simulations
Conduct Interventions
- Ablation Studies
- Generalization Tests
- Counterfactuals
Log Comprehensive Data
Phase 4: Analysis and Synthesis
Detailed Evaluation and Analysis
Compute Outcome Metrics
Compute Process Metrics
Analyze Communication
Perform Causal and Interpretability Analysis
Synthesis of Findings
Analyze Results
- Scaling
- Robustness
- Human comparison
Draw Conclusions on Emergent ToM Capabilities
Discuss scientific, ethical, and safety implications

Figure 1. The main Phases and Steps of the Proposed Multi-Agent Simulation Framework for TOM

2. Background and Related Work

2.1 ToM in cognitive science and developmental psychology

The modern research program on Theory of Mind (ToM) originates in Premack and Woodruff's seminal question, "Does the chimpanzee have a theory of mind?", which stimulated the study of mental state attribution in non-human agents [1]. In parallel, extensive research in developmental psychology has documented the trajectory of human ToM capabilities, exemplified by a child's comprehension of false beliefs around preschool age, classically assessed with the Sally-Anne test [7]. Persistent debates in cognitive science address whether human ToM is powered by a specialized innate component, a learned theory-building process, or a simulation of others' minds. These foundational insights inform the evaluation of artificial intelligence. Validating claims of artificial ToM requires grounding them in rigorous tasks and criteria analogous to those used in human studies. Furthermore, developmental delays in ToM, such as those associated with comparative primate studies suggest the involvement of specific cognitive mechanisms [8]. This raises the question of whether an artificial system, lacking analogous evolutionary and developmental structures, can attain similar abilities or is destined to fail in comparable ways.

2.2 Computational models of ToM and belief modeling

These psychological inquiries encouraged efforts to formalize the mechanisms of mental state attribution within computational frameworks. Influential approaches in Bayesian cognitive science treat ToM as a form of probabilistic inference, where an agent maintains a model of others' mental states and updates it based on observations, as seen in rational speech-act models of communication [9]. For instance, an agent might infer another's goal or knowledge by observing its behavior. Other research traditions have explored logic-based systems like dynamic epistemic logic to formally capture knowledge updates, as well as recursive agent models such as the Interactive Partially Observable Markov Decision Process (I-POMDP) formulation, which extends planning to multi-agent belief reasoning by explicitly incorporating nested beliefs into an agent's decision model [10–12]. While these computational models provide normative blueprints for ToM, they typically involve hand-crafted representations or operate in simplified settings, unlike the large, inductively learned models studied in this work. In order to address different aspects of ToM, our framework therefore includes parameterized task families, each procedurally generated to prevent data contamination and designed to make a specific order of belief reasoning necessary for success (Table 1).

Task Family	Primary ToM Target	Core Mechanic	Key Perturbation
False-Belief and Perspective-Taking	First-Order Belief Tracking	Asymmetric observation of object movement or visual	**Knowledge Swaps:** Exchanging agents' private observation histories while

		occlusion in an interactive world.	keeping surface text identical.
Deception and Signaling Games	Second-Order Strategic Reasoning	Misaligned payoffs in a cheap-talk game where a sender's message influences a receiver's action.	**Payoff Matrix Inversion:** Flipping incentives from cooperative to competitive to test strategic adaptation.
Coordination and Negotiation	Belief-Guided Joint Planning	Partially observable environments (e.g. mazes, resource allocation) requiring common ground formation.	**Partner Heterogeneity:** Introducing partners with novel or out-of-distribution communication styles.
Communication-Constrained Settings	Pragmatic and Belief-Sensitive Communication	Imposing costs or bandwidth limits on messages, forcing efficient information transfer.	**Bandwidth Ablation:** Systematically varying message length limits to measure pragmatic compression.

Table 1: Overview of ToM Task Families and Their Epistemic Demands

2.3 Emergence in large-scale models

The advent of such large-scale models has prompted claims of emergent abilities, which are capabilities that appear to surface once models reach a certain threshold of complexity or training size. Some researchers report that qualitatively new skills, including social reasoning, manifest suddenly as model size increases. An alternative perspective contends that many such instances of emergence are continuous performance gains misconstrued as discrete jumps due to coarse evaluation functions, dataset composition, or the long tail of task difficulty [13]. This view posits that apparent emergence is an artifact of nonlinear metrics or naive extrapolations of scaling curves. We treat emergence as a falsifiable hypothesis about patterns in our data, seeking evidence that ToM-relevant competence exceeds what naive scaling expectations would predict. Indeed, some scaling-law studies have found smooth, power-law improvements rather than strict discontinuities, suggesting many so-called emergent behaviors might simply require more fine-grained analysis. Nonetheless, the observation that certain abilities, like solving specific logic puzzles, appear to manifest only beyond a large parameter threshold keeps the discussion open. This stance shapes our experiments and their interpretation, namely any reported emergence is supported by statistical evidence beyond a simple parameter-performance curve and validated through replication and perturbation tests.

2.4 Multi-agent simulation for studying social cognition

Investigating such complex cognitive phenomena necessitates robust experimental paradigms. Interactive multi-agent simulations provide a promising opportunity to study social intelligence, creating dynamic scenarios where agents with partial information must interact over time [14,15]. Unlike static question-answer tests, these environments enable the study of communication, cooperation, deception, and the formation of shared conventions. Prior work has demonstrated that multi-agent environments can encourage emergent behaviors like tool use and novel communication protocols.

In the case of ToM research specifically, simulations offer a very important advantage, namely the ground-truth mental states of agents are known to the experimenter, allowing for objective measurement of whether an agent is successfully tracking what others know or believe [16]. This approach inspired our framework, which embeds models in environments designed to require belief reasoning, leading to reproducible data on their social cognitive performance. The use of multi-agent games also permits repeated trials under slightly varied conditions, which is extremely important for distinguishing consistent ToM competence from fortuitous guesses or overfitting [17].

2.5 The strengths and pitfalls of existing ToM evaluations for LLMs

The controlled, dynamic nature of multi-agent simulations stands in contrast to many existing ToM evaluations for large language models (LLMs) [17–19]. The question of whether LLMs possess ToM has led to a wave of evaluation attempts, often using classic false-belief stories or social vignettes. Although prompting techniques like chain-of-thought reasoning can elicit seemingly correct answers, these successes are difficult to interpret. An LLM might arrive at the right answer by leveraging superficial patterns or by recalling similar examples from its training data [18,19].

Reports of advanced ToM in certain models have been contested by findings that slight rephrasings or novel scenarios degrade performance, suggesting a fragile grasp of others' mental states. Existing evaluations are frequently one-off and brittle, lacking controlled difficulty or safeguards against data contamination. This methodological heterogeneity complicates efforts to ascertain whether a given model genuinely possesses ToM-like abilities or is merely exploiting shortcuts. These limitations motivate our controlled, simulation-based evaluation, which aims to provide stronger, auditable evidence regarding belief reasoning in AI. By engaging models in interactive, multi-turn tasks with verifiable belief dynamics, while also ruling out data contamination and providing robust analysis protocols, we seek to obtain a clearer picture of their social reasoning capabilities [20].

3. Problem Formulation

3.1 Operationalizing ToM (first-order, higher-order, recursive belief states)

This section establishes a formal framework for measuring Theory of Mind (ToM) in artificial agents, specifically those instantiated as policy wrappers for foundation models. The framework specifies the construct, the setting for its elicitation, and the conditions for interpreting measurements as genuine belief reasoning rather than as artifacts of surface pattern matching. This formulation aligns with the multi-agent simulation program where belief-dependent incentives, controlled partial observability, and instrumented communication render mental-state inference causally necessary for success. The entire approach is anchored by several foundational commitments. ToM is conceptualized as a graded, operational construct, distinct from a metaphysical attribute. Emergence is defined by robust, out-of-sample regularities, not by rhetorical discontinuities [19]. Valid measurement must integrate behavioral outcomes with underlying processes. Any claims must be falsifiable through perturbations that selectively impair belief-based solutions while leaving heuristic shortcuts intact. These principles ground the formalism and ensure its scientific rigor.

The formal specification begins by situating agents within a partially observable environment, modeled as a stochastic game, where each possesses unique observations and knowledge. An agent’s first-order belief is its internal estimate of the world's state, or task-relevant hidden information, given its observations. A second-order belief represents an agent’s model of another agent’s belief about the world. This recursion extends to higher-order beliefs, such as what one agent thinks another knows about its own knowledge. While a fully capable agent would theoretically maintain a nested model of others' mental states, the intractability of representing infinite recursive beliefs necessitates a functional approach. We therefore design tasks where optimal actions are contingent upon tracking the beliefs of others. An agent that consistently succeeds in these tasks is inferred to be leveraging the necessary belief-reasoning capabilities [8].

The methodology hinges on embedding carefully designed epistemic tests within the environment. These tests create situations where two scenarios are physically identical but differ in what another agent knows. An agent lacking ToM would perceive and act within these scenarios identically. A ToM-capable agent, however, would behave differently in response to the distinct epistemic conditions. We therefore consider ToM to be operationally present when an agent’s policy is sensitive to others’ mental states. This sensitivity is confirmed if the agent's behavior changes precisely in situations where another’s beliefs differ, even when the observable state of the world remains constant.

We adopt a minimal but expressive mathematical setting, namely a partially observable stochastic game with communication, because it canonically separates physical state uncertainty from epistemic state uncertainty and because it allows the faithful encoding of "who" knows "what", and "when". Let G denote a game with finite or countable horizon T, a set of agents $A = \{1, \dots, n\}$ a hidden world state $s_t \in S$ at time t, and, for each agent i, an observation function $O_i \colon S \times \varepsilon \to O_i$ mapping from the world state and recent events ε to a private observation O_i^t. Agents choose actions $a_i^t \in A_i$ and may emit messages $m_i^t \in M_i$ according to a communication protocol Π that specifies turn-taking, admissible content, and grounding conventions. The environment evolves via a transition kernel $P(s_{t+1} | s_t, a_1^t, \dots, a_n^t)$. Rewards r_i^t are assigned according to payoff functions that we intentionally couple to belief-relevant latent variables in order to manage the measurement of ToM.

In such a game, an agent's epistemic state is a distribution over latent variables given its observation–message history. We write $h_i^t = \left(o\frac{<t}{i}, m^{<t}, a^{<t}\right)$ for the history available to agent i at time t under the protocol Π. The agent's first-order belief about the world is a posterior $b_i^t(s) = Pr\left(s_t = s \middle| h_i^t\right)$. A Theory-of-Mind-capable agent maintains, explicitly or implicitly, higher-order beliefs about other agents' first-order beliefs, and potentially about their higher-order beliefs in turn. Let $b_{i \to j}^t$ denote i's belief over j's first-order world belief $b_{i \to j}^t(b) = Pr\left(b_j^t = b \middle| h_i^t\right)$. More generally, we define a recursive tower of beliefs $B_i^{(k),t}$ of order k, where $B_i^{(1),t} \equiv b_i^t$ and $B_i^{(2),t} \equiv b_{i \to j}^t$ for some $j \neq i$. Higher orders follow by induction, with $B_i^{(k+1),t}$ a measure over others' $B_j^{(k),t}$. Because explicit maintenance of full recursive distributions is intractable, we measure competence functionally, namely does the policy π_i select actions and utterances whose expected utility cannot be achieved without tracking certain beliefs to a specified order under the game's identifiability structure? The problem formulation therefore relies on tasks in which action values $Q_i\left(h_i^t, a\right)$ are provably sensitive to the content of $B_i^{(k),t}$ even after conditioning on $B_i^{(l),t}$for $l < k$.

In order to avoid confounds from purely world-state prediction, we engineer information asymmetries and intervention points where two histories h_i^t and h'^t_i induce the same first-order world posterior b_i^t but differ in the distribution over another agent's beliefs $b_{i \to j}^t$, and where the optimal a differs across h_i^t and h'^t_i. In such "epistemic disentanglement" regimes, performance demands sensitivity to others' beliefs. Communication protocols introduce pragmatic pressure, namely in cooperative tasks, a message is informative if and only if it reduces the entropy of the receiver's posterior over task-relevant variables. In competitive tasks, beneficial deception requires inducing a targeted misalignment between the receiver's posterior and the ground truth. The formal problem is therefore to determine, from behavior and process traces, whether a model wrapped as π_i leverages recursive belief information when such leverage is required by the incentive structure. This formulation is consistent with the article's stance that ToM is graded and operational and with its insistence on multi-turn interaction under partial observability to produce belief-sensitive behavior. These scenarios give us built-in controls for testing ToM. Consequently, we can systematically flip who knows what in otherwise identical trials and find out if the agent correspondingly flips its behavior.

3.2 Hypotheses and falsifiable predictions

Our formulation produces concrete, falsifiable predictions regarding the manifestation of ToM-like abilities. We hypothesized that first-order belief reasoning would be attainable for current large language models [19], particularly with appropriate interaction protocols, whereas reliably handling second-order beliefs would present a significant challenge. This tiered-difficulty hypothesis predicted that performance would decline sharply with each increase in required reasoning, with second-order tasks producing results at or near chance levels. We also advanced a scaling hypothesis, positing that larger models would substantially outperform smaller ones on belief-critical tasks, suggesting that social reasoning capabilities emerge at scale, albeit with diminishing returns at higher orders of inference [21].

Further predictions concerned the experimental conditions. We anticipated that forcing agents to communicate and coordinate would reveal more pronounced performance differences, with failures emerging in tasks requiring subtle pragmatic inference or deception [22,23]. A scaffolding hypothesis proposed that agents equipped with support mechanisms, such as chain-of-thought reasoning or memory tools, would achieve higher ToM scores than those relying solely on base model responses. In order to assess whether observed competence was intrinsic, a robustness hypothesis stipulated that genuine ToM should remain stable across variations like paraphrased scenarios or randomized partner behaviors. Consequently, a significant performance drop under such perturbations would indicate a reliance on brittle shortcuts. All hypotheses were pre-registered where possible in order to ensure our analysis rigorously sought to falsify these predictions .

3.3 Threat models and need for valid ToM measurements

Our measurement approach was designed to guard against several ways an AI system might simulate ToM without genuine ability [24]. A primary threat is shortcut exploitation, where a model leverages pretend cues memorized from training data. In order to counter this, our task generators produce numerous scenario variations with different surface details, preventing simple pattern recognition. Another threat is overfitting to a partner's behavior, where an agent learns a fixed response script that works by coincidence. We mitigate this by randomizing partner policies and roles, forcing the agent to adapt rather than rely on static assumptions. We also prevent information leakage by ensuring each agent's observations are strictly private, and that controlled communication channels do not inadvertently reveal hidden knowledge.

These safeguards inform our core design decisions. The principle of partial observability ensures agents never have direct access to each other's private information [25]. Tasks are built around informative interactivity, requiring information exchange or observation for success. The methodology includes counterfactual tests, such as swapping which agent holds a very important piece of knowledge, in order to verify that behavior changes appropriately. In addition, role symmetry, achieved by having agents swap roles across trials, ensures that strategies are not tied to superficial identity cues [18]. We also address the risk of training data contamination by using novel scenarios and studying any exceptional performance on tasks that resemble known puzzles [17]. The evaluation environment itself is used to log all information an agent receives, enabling audits that confirm no hidden channel provides a shortcut to the solution. These combined measures ensure that high performance is attributable to effective mental state inference.

3.4 Construct, internal, and external validity criteria

We have specified precise criteria to ensure our measurements of ToM are valid and meaningful. Construct validity, which confirms that our tasks capture the essence of ToM, was established by aligning them with classic psychological paradigms like false-belief tests. The tasks were designed so that success requires belief reasoning, precluding solutions based on simple heuristics or lucky guesses.

Internal validity, the basis for drawing causal conclusions, was addressed through tight experimental controls [26]. We employed ablation studies to observe whether removing a key component, such as memory or communication, degrades performance in a manner consistent with the loss of ToM reasoning. We further enhanced internal validity through pre-registered analyses, appropriate statistical models, and adversarial trials that test whether apparent skills are robust under stress. These practices collectively separate genuine abilities from experimental artifacts.

External validity concerns the generalizability of our findings beyond the specific simulation [27]. While our multi-agent games are abstractions, they capture core dynamics of belief reasoning. We acknowledge their limitations, namely success in this controlled world does not imply a full understanding of human beliefs in open-world settings, which involve richer social cues. In discussing our results, we carefully differentiate between abilities demonstrated within our tasks and the broader complexities of human social cognition. We are also explicit that passing our tests does not render an AI safe or socially aligned for deployment. By clearly defining these validity boundaries, we provide a solid foundation for interpreting the presence or absence of ToM-like behavior in the agents we study.

4. Multi-Agent Simulation Framework

4.1 Environment design and task families

Our evaluation platform is a custom multi-agent simulation environment, specifically designed as a partially observable, turn-based world to produce behaviors dependent on Theory of Mind (ToM). The tasks are configured so that success hinges upon an agent’s ability to track what other agents have or have not observed [2,3]. Within this shared environment, multiple agents act based on their own private observations. In order to probe distinct aspects of belief reasoning, we have developed several families of tasks.

The first family adapts classic false-belief scenarios into interactive games. In a typical setup, one agent hides an object while another is not observing, and the second agent must then locate it. The first agent's capacity to predict the other's mistaken belief is tested over multiple turns as agents move or communicate, requiring a consistent application of ToM. Another set of tasks explores deception and bluffing, drawing from cheap-talk and costly-signaling games. Here, a "sender" agent, who knows a particular truth, benefits if the "receiver" holds a false belief. The sender's success requires modeling how the receiver will update its beliefs based on a potentially deceptive message. Conversely, the receiver must reason about the sender's motives, probing the agents' capacity for intentional deception and its detection.

A third category involves cooperative and competitive coordination tasks. In cooperative scenarios, agents with partial information must infer each other's knowledge to achieve a common goal, such as meeting in a maze. In competitive variants, an agent might exploit another's ignorance to gain an advantage. These situations demand reasoning about others' knowledge and intentions in the context of joint action. We study the robustness of these behaviors in communication-constrained settings. By limiting the number or increasing the cost of messages, we test whether agents can convey critical information efficiently. These constraints ensure that emergent communication strategies are genuinely informative and not merely exhaustive exchanges of data. Beyond these core families, the environment incorporates further nuances to ensure robust evaluation. We introduce irrelevant distractor objects to prevent agents from focusing on a single variable and vary tasks between one-shot and repeated encounters to observe if agents learn and adapt conventions over time.

4.2 Agent architecture and scaffolding

Each simulated agent is implemented as a policy wrapper around a pretrained foundation model, which functions as the agent's cognitive core. This wrapper manages the flow of information, feeding the model appropriate inputs and translating its outputs into actions or messages. The architecture comprises several key components. At each turn, the wrapper constructs a detailed prompt that provides role context, including the agent’s identity, a summary of recent events, and output format guidelines. This structured context grounds the model in the simulation’s state and its designated role. The architecture also includes a memory buffer for recent observations and actions. Extremely important, an agent's internal state features a belief model for tracking what it presumes another agent knows.

This model is populated exclusively through inference based on observable actions and communication, not through privileged access to the simulation's ground truth. This methodological decisions ensure that successful belief reasoning originates from the model’s learned capabilities rather than from externally supplied information. In order to facilitate complex decision-making, the wrapper provides an optional planning scratchpad where the model can perform private, intermediate reasoning, akin to a chain of thought, before committing to an action. Agents may also call simple tools to query hypothetical outcomes. This scaffolding enhances performance on multi-step problems and allows us to distinguish the base model's unaided capabilities from what it can achieve with external support by toggling these aids during experiments.

4.3 Communication protocols

Communication between agents is mediated by a structured protocol governing textual messages, ensuring that all content can be analyzed for its grounding and informativeness [28]. Each message must adhere to a specific format, including a content string and optional references, or grounding handles, that tie a statement to a specific observation. This system prevents agents from making claims about facts they could not have perceived. The protocol is further defined by turn-taking rules and an infrastructure that serializes all interactions into append-only logs with cryptographic checksums, creating an evidential trail for all claims.

While the protocol provides structure, it is flexible enough to allow for the emergence of communicative conventions and pragmatic inference. Over repeated interactions, agents can develop nicknames for landmarks or code words for routine actions, particularly when communication is restricted or costly. We detect the formation of such conventions by observing a decrease in message surprisal over time coupled with stable or improved task performance. Our analysis extends beyond literal content to pragmatics, examining how agents infer meaning from context. A telling silence, a well-timed interjection in an interruptible regime, or the frequency of confirmation queries all serve as signals that reflect an agent’s confidence and shared understanding. These pragmatic and efficiency metrics ensure that success arises from concise, cooperative information exchange established in belief reasoning.

4.4 Reward structures and curricula

The environment’s reward schemes are carefully crafted to incentivize belief-aware strategies [28]. The primary reward is goal-based, such as a positive value for successfully completing the task. The tasks are designed so that achieving this goal inherently requires correct belief reasoning. In order to guide learning, we sometimes include auxiliary shaping rewards, for instance, a small bonus for sending a helpful, truthful message. These shaping rewards are carefully restricted to epistemically meaningful quantities, like the reduction of uncertainty in a partner’s inferred belief state and are gradually strengthened as training progresses so that agent competence ultimately relies on the primary task reward. Our training regimen combines self-play and curriculum learning.

In self-play, agents from a continually refreshed pool are paired, and their roles are regularly swapped to prevent the memorization of a specific partner's behavior and encourage the development of generalizable strategies. The curriculum begins with simple scenarios and progressively increases in difficulty, introducing more subtle belief states or deeper levels of recursive reasoning to produce higher-order ToM. In order to promote generalization, we also employ extensive domain randomization. Superficial elements of the tasks, such as room layouts, object names, and agent identifiers, are varied widely across episodes. This randomization, combined with a default policy of resetting the state after each episode, ensures that learned behaviors are a response to the latent epistemic structure of the tasks rather than an overfit to surface features. These practices collectively ensure that any emergent ToM-like behavior is a robust response to the task's fundamental demands.

5. Models and Training/Inference Regimes

5.1 Foundation models evaluated (sizes, training data, modalities)

Our evaluation encompassed a spectrum of state-of-the-art foundation models, which served as the cognitive cores for our agents [2,3]. This selection included large language models of varying scales, from several hundred million to tens of billions of parameters, trained on diverse textual corpora [17,19]. The models represented different architectural families, including those based on the GPT architecture and other Transformer variants. Our collection ranged from models trained exclusively on text to those with limited multimodal pre-training in vision and language. Although our primary focus was on text-based reasoning to align with the textual nature of our tasks, we considered whether multimodal models might offer intrinsic advantages, given that human Theory of Mind often integrates visual cues. All models were initially employed in their pre-trained state without task-specific supervised fine-tuning to assess their zero-shot capabilities on these novel interactive challenges.

The model pool included a 350-million-parameter Transformer LM [29] pre-trained on internet text, a 6-billion-parameter model with a similar training regimen but greater capacity, a large 70-billion-parameter model featuring extensive instruction-tuning, along with a representative multimodal model capable of processing image-like inputs. For our text-based simulations, the multimodal model functioned equivalently to its text-only counterparts. We verified to the best of our knowledge that each model possessed general world knowledge and language proficiency but had not been specifically trained on our tasks or on simplistic Theory of Mind puzzles[30] .

5.2 Prompting strategies (zero-shot, few-shot, role prompting, tool-augmented)

We systematically varied the prompting strategies used to guide model behavior [7,31]. In the zero-shot configuration, the agent received only the minimal context of its observations and the game state in natural language, compelling it to act based on its pre-trained knowledge alone. We also implemented few-shot prompting, where the model was provided with one or more complete example episodes demonstrating belief-aware behavior before it attempted a new task. This method helps the model recognize the required pattern of reasoning about others. Furthermore, we emphasized role prompting, which explicitly informed the model of its identity and objectives as described in Section 4.2. A prefix such as, "You are a hider. Your partner is a seeker. You want them to have a false belief about where the treasure is", effectively primed the model for a strategy of deceptive communication.

For certain conditions, we augmented the prompts with tools or a scratchpad, allowing the model to "think" silently before responding or to call external functions like memory retrieval. These variations enabled us to test how performance gains scaled with instructional quality versus the model’s intrinsic capabilities. We found that few-shot prompting frequently enhanced performance on first-order tasks by illustrating the concept of information sharing or withholding. Tool-augmented prompting proved especially beneficial in scenarios demanding multi-step planning, where the model could outline a deceptive strategy in its scratchpad before execution. The design of our prompts required a careful balance. On one hand, prompts must define roles, goals, and the output schema to ensure well-formed actions and messages. On the other hand, over-engineered prompts could inadvertently leak solutions, for instance by stating "remember, the other agent hasn’t seen X". We therefore iterated carefully on prompt design to ensure the prompts guided behavior without revealing answers. The resulting strategies span from minimal role hints to detailed, step-by-step formats, and we report results across this entire spectrum.

5.3 Fine-tuning and reinforcement learning settings

While much of our analysis treats the foundation models as static, we have also studied parameter-updating approaches in order to determine if the models could learn Theory of Mind through direct interactive experience [6,32]. In a supervised fine-tuning framework, we used interaction transcripts from our simulation, some of which included demonstrations from an oracle agent exhibiting correct belief reasoning. The language model was then fine-tuned to better predict the correct outputs in these scenarios, effectively teaching it through gradient descent on example episodes. Concurrently, we experimented with reinforcement learning (RL) [33], where task rewards provided direct feedback to optimize the agents' policies. In these RL sessions, two agent instances played numerous episodes, and the model's parameters were updated using policy gradient methods to maximize expected reward.

The fine-tuning approach passes on explicit knowledge of Theory of Mind tasks, namely if a model initially fails to grasp the significance of another agent's perspective, the fine-tuning examples make this significant. The RL approach, conversely, allows for the discovery of strategies beyond the provided examples, as the agent is free to explore any behavior and receive feedback [33]. However, RL is susceptible to finding local optima. For example, an agent might discover a simple exploit that wins the game without genuinely developing Theory of Mind, a possibility our environment design sought to minimize. We observed that supervised fine-tuning on a small set of demonstration episodes yielded noticeable improvements in immediate performance, particularly in making the model's communications more relevant to its partner’s knowledge state, though these improvements eventually plateaued. With RL training, agents demonstrated further gains on some tasks after many iterations, indicating they can learn superior policies through self-play. Interestingly, some RL-trained agents developed novel conventions distinct from those we might have hand-coded, emphasizing the creative potential of direct goal optimization. We also noted that RL sometimes produced poor behaviors, such as overfitting to specific training scenarios, highlighting that generalization remains a challenge without sufficiently diverse training conditions.

5.4 Decoding, temperature control, and self-consistency

The method by which a model generates its response significantly influences its performance in multi-agent settings. We examined different decoding strategies for the language model outputs[3,23]. For example, a higher temperature encourages more exploratory and diverse responses, which can produce creative solutions but also risks incoherence. A lower temperature yields more deterministic and conservative responses. In order to ensure fair comparisons, we generally kept decoding settings constant but tested the extremes to observe their effects on tasks requiring subtle reasoning. A moderate, slightly stochastic decoding process proved most effective. Overly deterministic generation could cause an agent to fall into a repetitive loop, while excessive randomness could lead to irrelevant outputs that violate the interaction protocol. In some cases, we also employed a self-consistency approach. This involved having the model internally generate multiple potential actions or messages and then select the final output through a voting mechanism or by choosing the option with the highest confidence. This technique can mitigate random errors by leveraging the model's own uncertainty estimates. If a model generated three potential moves, two of which originated from a correct inference about another agent's false belief, this method would favor one of those two, thereby amplifying consistent reasoning.

Decoding choices shape agent behavior as profoundly as parameter count. Parameters such as temperature, nucleus thresholds, repetition penalties, and stopping criteria all influence whether an agent's dialogue is terse or verbose, literal or inventive. We systematically fixed or swept these parameters as interventions, recorded them in the experimental manifest, and included them as covariates in our analysis. This practice prevents agents with "chattier" decoders from gaining an unfair advantage in human-rated metrics and upholds the rigorous governance we established for our experimental configuration. Careful tuning of the generation process was very important for reliable performance, confirming that an apparent cognitive failure could come from the generation process itself rather than from a fundamental model limitation.

5.5 Safety filters and chain-of-thought handling policies

The inclusion of scenarios involving deception necessitated a deliberate approach to managing potential conflicts with AI safety and alignment measures [34,35]. Certain large language models have built-in safety filters that discourage the production of manipulative or untruthful content [17–19]. In our context, however, misleading another agent is a legitimate strategic component of a deceptive role. We therefore carefully configured the models to understand that in-game deception was a permissible part of the simulation. This involved phrasing prompts to clarify that the agent was playing a character in a fictional context, framing any "lie" as an action within that context rather than a violation of the model's instruction to be truthful. For instance, we added clarifications such as: ("You are role-playing in a game scenario, statements in the game are not real assertions to the user."). We ensured that the model's internal reasoning, or chain-of-thought, was kept private and that all its public outputs were appropriate. We separated the model's private reasoning from its public messages, either by using special tokens or by splitting generation into two phases by first generating thoughts, then generating the observable message. Log verification confirmed that no unintended information was leaked, preventing an agent from accidentally revealing its private notes to its partner.

Strategic deception and persuasion are scientific targets that also raise ethical questions [36]. In our study, all agents were AI, with no humans involved in gameplay, which eliminated the risk of human deception. However, when considering deployment, it is extremely important to ensure that an AI capable of deception in thesetasks does not engage in dishonesty outside of carefully defined roles. Our experiments, by highlighting how and when an AI might choose to lie for strategic advantage, may inform the design of future safety filters. One might implement a policy that prevents deceptive content unless a specific flag indicates a permitted simulation scenario. We have balanced safety considerations with game realism by explicitly delineating the boundaries of deception in prompts and by technically isolating internal reasoning from external output. This approach allowed our agents to fully engage in behaviors like bluffing where appropriate, without violating the principles of safe model deployment. This process also provides valuable insight into how future AI systems might be gated or contextualized to use their Theory of Mind abilities in beneficial, rather than harmful, ways.

6. Evaluation Protocol

6.1 Datasets and scenario generation

The evaluation protocol is designed to determine, with the highest practical degree of internal and construct validity, whether an agent exhibits Theory of Mind (ToM)-relevant competence within our multi-agent simulations. This objective necessitates carefully curated scenarios and comparisons. Scenarios are generated through parameterized programs that vary the surface realization of instances while preserving the latent epistemic structure essential for ToM. For each family of tasks, we have designed a set of scenario templates, which function as probabilistic grammars that instantiate diverse settings. For instance, false-belief tasks might involve objects hidden in boxes, characters moving between locations, or secrets concerning identities. The generators randomize details such as object names, spatial arrangements, and event timings, ensuring no two episodes are identical in their phrasing or layout, even while they structurally assess the same belief-reasoning construct.

In order to target specific hypotheses, we have partitioned these scenarios into distinct evaluation sets. One set might contain straightforward false-belief tests to establish a baseline, while another could introduce additional distractors or require second-order reasoning for success. These distinct scenario families enable us to study important questions, such as whether an agent that passes simple false-belief tests can generalize to more difficult ones, and how its performance compares to agents employing alternative strategies. Each scenario is labeled with metadata that specifies which agents possess privileged information, the turn at which a false belief might arise, and the optimal strategy, such as communication or deception. This metadata is very important for analysis, as it allows us to stratify results and distinguish episodes that genuinely require second-order ToM from those solvable with first-order reasoning, thereby ensuring that our summary statistics are coherent and meaningful.

6.2 Baselines, random scriptable agents, symbolic ToM models, human benchmarks

The interpretation of an agent's performance requires comparisons against a comprehensive suite of baselines, which anchors our claims across the full spectrum from chance to ideal performance [18,37]. The most fundamental baseline is a random policy agent that makes moves or sends messages uniformly at random from the set of valid options. This agent represents the performance of a system with no task understanding and establishes a lower bound, which is often near zero percent success in complex tasks. Afterwards, we implemented several simple, rule-based heuristic agents designed without explicit ToM. A heuristic seeker, for example, might always search the last location where it observed the hider, ignoring the epistemic state of the hider. Similarly, a heuristic communicator might always be truthful or deceptive depending on the goal. These agents provide mid-level baselines, namely if our learning agents fail to outperform a naive rule, it indicates they have not acquired the intended reasoning. In order to define the upper bound of achievable performance, we include symbolic or optimal ToM models where feasible.

In the case of certain tasks in constrained state spaces, we can design a dynamic Bayesian network that explicitly tracks all agents' knowledge and selects optimal moves [38]. In other cases, we use game-theoretic solutions. These "oracle" agents, representing ideal reasoners, provide a ceiling against which to measure the learned agents' proficiency. We have used human benchmarks to contextualize performance and have referenced established results from psychology, such as the typical success rates of children versus adults on specific ToM problems. In select cases, one can conduct tests with human participants, providing them with the same information as an agent to confirm that the tasks were indeed solvable by a mind with ToM. As humans bring a lifetime of social intuition, they provide a "gold standard" for performance. By comparing our agents against this spectrum of baselines, we ensure that claims of ToM competence are rigorously grounded. An agent performing near the heuristic baseline likely uses a simple strategy, whereas one approaching the symbolic ideal provides strong evidence of nontrivial belief reasoning.

6.3 Ablation studies (memory off, comms off, planning off, role swaps)

Ablation studies form the core of our causal inference, allowing us to attribute observed behaviors to specific architectural components [39,40]. We conducted a series of these studies to pinpoint which elements of our agent design truly contribute to ToM-like performance. In one study, we disabled the agent's memory module and belief tracking, forcing it to rely solely on the current observation. Performance on tasks requiring memory of who saw what, such as false-belief scenarios, dropped dramatically. Agents frequently acted as if others shared their observations, becoming effectively mind-blind. Another intervention prohibited communication, which revealed the extent to which success depended on explicit information exchange. In many cooperative tasks, agents that had previously succeeded with ease became unable to solve the problem, highlighting their reliance on communication over independent inference. We have also disabled the chain-of-thought planning scratchpad, compelling the model to produce an action immediately. This change particularly degraded performance on higher-order tasks, as agents resorted to greedy, one-step decisions and failed to anticipate the future epistemic states of their partners. A final ablation involved swapping agent roles to test for generality. Agents trained with role randomization adapted seamlessly, whereas those trained asymmetrically often failed when their role was reversed, indicating their learned strategies were not robust.

These interventions, where each agent serves as its own control, allow us to isolate causal relationships. The resulting performance deltas, analyzed with cluster-robust intervals, attribute variance to the ablated component. Our findings confirmed pre-registered hypotheses, for instance, second-order reasoning was far more reliant on the chain-of-thought feature, supporting the idea that deeper reasoning benefits from internal deliberation. We have also supplemented these quantitative findings with qualitative reviews of transcripts. An agent without memory, for example, would often produce contradictory information, while an agent without communication sometimes attempted to signal through its actions, revealing interesting emergent workarounds. These qualitative differences reinforce the conclusion that each component played a distinct and critical role.

6.4 Generalization tests (novel tasks, out-of-distribution agents, few-shot transfer)

Generalization has been assessed along three primary axes, namely scenario novelty, partner novelty, and training regime transfer. In order to test for scenario novelty, we have evaluated agents on variant tasks that were not part of their training curriculum. For instance, an agent trained on two-room hiding games might be tested on a three-room version or a conceptually similar card game. We have found that first-order skills, like tracking who knows what, often transferred well to superficially new contexts, whereas more complex strategies like multi-turn deception proved more weak. For partner novelty, we have paired our agents with partners exhibiting behaviors outside the training distribution, such as a scripted agent with quirky, random actions or an agent from a different model family. These tests reveal whether an agent's ToM is robust or over-tuned to its self-play training partners. Results showed a performance dip when facing unfamiliar partners, indicating that agents had developed specific expectations that did not always hold. Agents trained with greater partner diversity, however, were more resilient.

We also assess few-shot transfer by providing an agent with a small number of examples to adapt to a new condition, such as shifting from a cooperative to a competitive game. We have observed limited but notable adaptability, particularly in larger models, which could often infer a new rule from a handful of demonstrations and adjust their behavior accordingly. A final test measures cross-episode generalization by correlating an agent's performance across structurally similar episodes. High-performing agents demonstrated strong consistency, reliably solving problems of the same reasoning type, whereas medium-performing agents were more erratic. Our generalization tests have indicated that the agents have learned abstract skills, but this generalization is not universal and remains bounded by their training experiences.

6.5 Cross-lingual and multimodal variants

In order to validate that the measured construct is authentically epistemic rather than purely linguistic, we have integrated cross-lingual and multimodal evaluations as a core component of our protocol. [39,41–43] In our cross-lingual experiments, we translated simulation narratives and agent communications into other languages understood by the models, such as French. This tests whether an agent's ability is based on conceptual understanding or merely on pattern-matching English-specific cues. In the case of multilingually trained models, performance remained strong after translation, with only slight drops attributable to data distribution differences. This outcome indicates that the agents grasped the situations at a conceptual level. For multimodal considerations, we conducted limited tests by supplementing textual descriptions with simple schematic images encoded textually. While full visual understanding was beyond our scope, these forward-looking experiments explore how ToM reasoning could extend to visual perspective-taking. Even rudimentary spatial information helped agents clarify perspectives more easily, a finding that aligns with the human reliance on visual cues for ToM tasks. These cross-lingual and multimodal evaluations are very important for demonstrating the generality of our findings. They confirm that the measured capability is not language-bound and lay the groundwork for future work incorporating richer perceptual inputs, such as first-person visual feeds for each agent.

6.6 Leakage controls and contamination checks

Given that foundation models are trained on vast, heterogeneous quantities, information leakage presents a significant challenge to experimental validity [30,41,44]. Our protocol implements controls at the generation, runtime, and analysis stages to ensure models do not covertly exploit information. During scenario generation, we avoid well-known common tests and produce multiple variants of each scenario in order to prevent recognition. At runtime, we strictly enforce observation restrictions, ensuring agents cannot reference information they have not perceived. The environment automatically flags any output containing privileged information, which we then mark for further analysis. All interactions are hashed and timestamped to support post-hoc audits for potential pre-training contamination. During analysis, we perform contamination checks by searching the model's known training data for key phrases or structures from our scenarios. When a scenario resembled a classic example like the Sally-Anne test, we paraphrased it, changed names, and added extra steps to eliminate one-to-one correspondence. This controlled approach provides confidence that our results are trustworthy. We are ensuring that positive findings are registered due to the model's genuine inferential capabilities, not from memorization or accidental information exposure, and that negative results are not due to avoidable flaws in our experimental design.

7. Metrics and Statistical Analysis

7.1 Outcome metrics, task success, belief inference accuracy, ToM tier scores

We first evaluate each agent's performance through its task outcomes. The most direct metric is the success rate for each scenario, which assesses whether the agent has achieved its payoff-relevant objective given the scenario's hidden state and payoff matrix [42]. Success is defined contextually, namely in cooperative search, it means jointly locating and retrieving an object within budget. In deception and signaling games, it requires the sender to induce a belief update that yields a target action from the receiver. In coordination problems, it is achieved when joint choices satisfy the payoff-dominant equilibrium. We compare these success rates to baseline policies to confirm that the agent performs above chance or simple heuristics. Recognizing that not all successes are equivalent, we focus on belief-critical decision points within each episode. These are junctures where the correct action depends on accurately understanding another agent's knowledge or beliefs. We measure the proportion of these critical points at which the agent selected the belief-sensitive optimal action. This approach sharpens construct validity by filtering out routine segments of an episode and directly tying performance to the agent's use of epistemic leverage.

We also directly evaluate belief inference accuracy by comparing each agent’s latent or reported beliefs against the simulator’s ground truth. In the case of policies that expose explicit belief reports, such as a probability distribution over world states, we measure alignment using proper scoring rules like the Brier score and cross-entropy. These scores are normalized by the entropy of the ground-truth distribution to enable difficulty-aware comparisons across tasks. When policies do not provide explicit belief states, we infer a behavior-implied belief by inverting an action-value model calibrated on the environment’s dynamics. An action that is optimal only under a specific belief about a partner implies a posterior over that partner’s beliefs. We then score this inferred belief against the ground truth using the same scoring rules. This dual approach credits implicit Theory of Mind (ToM) while guarding against persuasive but unfaithful verbal rationales.

We have computed ToM tier scores to summarize performance. Each episode possesses a hidden construct signature that identifies the order of belief required for optimal decisions. We label each decision point by tier, such as first-order for tracking a partner’s beliefs about the world or second-order for tracking a partner’s beliefs about one’s own beliefs. An agent’s score for an episode is the proportion of optimally resolved decision points at each tier. The final ToM tier score is a weighted sum across tiers, with weights reflecting both the tier order and the decision point’s importance to the expected return. This weighting ensures that sparse but decisive second-order points are not overshadowed by numerous low-stakes first-order points. This tiered evaluation reveals the depth at which an agent’s recursive reasoning fails, offering a clear profile of its ToM limitations. In order to standardize comparisons, we normalize scores between a naive baseline and an ideal reasoner’s performance, reporting results as a percentage of this range.

7.2 Process metrics, justification quality, belief alignment over time

Outcome metrics are complemented by process evidence, as ToM fundamentally involves representing and updating others’ mental states [41–43]. We have used agent outputs to capture both verbal justifications and nonverbal signals of belief tracking over time. Justification quality is evaluated from message transcripts by human raters who are blinded to the experimental condition and model identity. Raters assess whether a justification correctly identifies the belief holder, cites relevant observations, and updates beliefs appropriately. They also judge if the rationale anticipates the interlocutor's interpretation of a message. We quantify inter-rater agreement using Krippendorff’s α for categorical items and an intraclass correlation coefficient for scalar scores. In the case of automatic analysis, we also compute the text similarity between agent justifications and ground-truth explanations, which are derivable from the known environment. These metrics verify that an agent’s reasoning is correct not just in its conclusion but also in its process. Belief alignment over time measures how closely an agent’s reported or behavior-implied beliefs track the ground truth as an episode unfolds. We compute a per-timepoint divergence, such as the Kullback–Leibler divergence, between the agent's estimate of another’s beliefs and the simulator-computed ground truth. We then summarize each episode with a discounted cumulative alignment score, which upweights early, belief-setting moves. A high alignment score indicates that an agent reached success through consistent perspective tracking rather than fragile heuristics. An agent with proficient ToM maintains near-zero divergence, whereas an agent that confuses perspectives exhibits spikes in this error measure.

We also track belief–action coherence by scoring the consistency of an action with the agent's previously reported beliefs [43]. An action that contradicts a stated belief under the optimal policy is penalized as incoherent. Analyzing these process measures allows us to identify successes achieved through luck or non-ToM heuristics, which typically manifest as high task success coupled with poor justification quality or low belief alignment. Conversely, high belief alignment preceding a failure due to an unpredictable event suggests an ability deficiency rather than a performance artifact. Belief alignment over time measures how closely an agent's reported or behavior-implied beliefs track the simulator's ground truth as the episode unfolds. Let b_t^i denote the simulator-computed belief distribution for agent i at time t given their observation history, and let $\hat{b}_t^{i\to j}$ denote the focal agent's estimate (reported or behavior-implied) of i's beliefs at time t. We compute a per-timepoint divergence $D_t = KL\left(b_t^i \parallel \hat{b}_t^{i\to j}\right)$ or its symmetrized Jensen–Shannon variant when beliefs are multimodal. We summarize each episode with a discounted cumulative alignment score $A = \sum_t \gamma^t \left(1 - norm(D_t)\right)$, where $norm(\cdot)$ maps divergences to [0,1] by reference to the divergence of a uniform posterior. The discount γ upweights early, belief-setting moves in coordination games and equalizes the contribution of long and short episodes. This alignment score is reported jointly with outcome metrics. Their correspondence helps distinguish agents that reach success through consistent perspective tracking from those that succeed via fragile heuristics, namely the importance of aligning process traces with veridical belief states made possible by logging observations, actions, and messages with epistemic tags and grounding handles, an aspect that was emphasized in our proposed design.

In simpler terms, belief alignment is asking the question "as the process goes on, does the agent's internal model of what others know remain accurate? If, at time, the partner should have a 70% chance of believing X (given what they've seen), does our agent also act as if the partner has ~70% confidence in X (say, by how it communicates or by what it expects the partner to do)?". We track a divergence or error measure at each step. A perfectly ToM-competent agent would keep this divergence near zero throughout, whereas an agent that sometimes forgets or confuses perspective would show spikes (like a high KL divergence at the point where it makes a mistake about what the other knows). We have also tracked belief action coherence by scoring the consistency of an action with the agent's previously reported beliefs. An action that contradicts a stated belief under the optimal policy is penalized as incoherent. Analyzing these process measures allows us to identify successes achieved through luck or non-ToM heuristics, which typically manifest as high task success coupled with poor justification quality or with low belief alignment. Conversely, high belief alignment preceding a failure due to an unpredictable event suggests an ability deficiency rather than a performance artifact.

7.3 Pragmatics and communication efficiency measures

Communication is the primary medium for revealing and manipulating beliefs [41–43]. Our efficiency measures therefore link message content to the receiver's epistemic state rather than to surface features like length. We have used pragmatic informativeness as expected posterior contraction for the receiver. For a message m emitted at time t to receiver r with pre-message posterior p_t over a task-relevant latent X, the informativeness is $I(m \rightarrow r) = \mathrm{E}_{p_t}[H(X|p_t) - H(X|p_{t+1})]$, where p_{t+1} is the receiver's posterior after updating on m under the protocol's grounding rules. We have estimated this from the simulator's reconstruction of r's posterior. In cooperative tasks, higher expected contraction is better, conditional on truthfulness. In competitive tasks, misinformative messages are rewarded only when payoff structure justifies strategic obfuscation, and we have caped rewards to prevent runaway miscalibration incentives as described in our reward-shaping policy. The explicit dependence of shaping terms on the receiver's posterior, rather than on ground truth alone, avoids smuggling in answers and aligns pragmatic scoring with the recipient's epistemic state.

In simpler terms, we measure how much each message reduces uncertainty for the other agent. If a message does not tell the partner anything they could not infer, it is low informativeness. If it greatly reduces their uncertainty (either by providing a clue in a cooperative game or by sending them down a wrong path in a deception game), it is high (pragmatically) informative. We also account for misinformation, namely a message that leads the partner to have false beliefs might be scored negatively in cooperative contexts (where it is just confusion) but can be part of the strategy in competitive ones (there, it might earn positive reward if it confers advantage, although within limits in order to avoid encouraging the model to hallucinate wildly beyond what the game incentives justify).

Efficiency also depends on channel constraints [41]. Bits-per-decision-point, the ratio of posterior contraction to message length, quantifies communicative economy. We have also scored time-sensitive pragmatics, for instance, in interruptible regimes, silence can signal deference or confidence. The use of costly listener check-backs, such as optional confirmations, offers another perspective on pragmatic competence. Overuse may signal pessimistic beliefs about a partner's comprehension, while underuse can indicate overconfidence. We have also measured convention alignment. In repeated interactions, agents often develop emergent conventions for communication. We have fit a dynamic language model to a partner pool's messages and compute the surprisal of each new message. Decreasing surprisal over time, paired with stable informativeness, signals the emergence of efficient conventions rather than routine collusion. We analyze this interchange alongside outcome metrics to distinguish healthy pragmatic adaptation from weak, non-generalizable codes.

7.4 Calibration, confidence elicitation, and over/under-confidence

Calibration measures the alignment between an agent's expressed confidence and its empirical accuracy [41,45]. We elicit confidence in two ways, either directly from policies that emit explicit posterior probabilities, or via a scalar self-report for categorical claims. In both cases, we compute metrics such as Expected Calibration Error (ECE) and the Brier decomposition. A reliability diagram with a slope of 1 and an intercept of 0 signifies perfect calibration. We evaluate calibration at belief-critical decision points, as it is most important for agents to act on well-calibrated uncertainty when the world is epistemically ambiguous. We also measure calibration-conditioned pragmatics by correlating confidence with communication choices. Well-calibrated agents should reserve expensive signals for high-leverage, high-uncertainty states. Our calibration suite quantifies these tendencies and links them to specific failure modes, such as sending a misleading message at the wrong time due to a misjudgment of another's knowledge.

Over and under-confidence are quantified as directional deviations between confidence and empirical accuracy. We define over-confidence as the signed error $OC = \frac{1}{N}\sum_n(c_n - a_n)$, where c_n is confidence and a_n is outcome accuracy for instance n. We complement this with threshold-conditioned measures, namely the false positive rate among high-confidence claims and the false negative rate among low-confidence claims. Because deception games incentivize selective misrepresentation, we distinguish epistemic miscalibration (confidence misaligned with accuracy) from strategic misrepresentation (confidently stating a belief known to be false to manipulate the partner). The latter is detected when the agent's internal state or prior belief report diverges from its public message in a direction that improves payoff under the partner's expected update model. One should note that over-confidence in false beliefs and under-confidence in correct inferences co-occur and that misalignment between self-reports and subsequent actions is a recurrent failure mode. Our proposed calibration suite quantifies these tendencies and ties them to pragmatic choices.

7.5 Reliability, test–retest, inter-rater, and inter-scenario consistency

Reliability ensures the stability of our measurements across minor perturbations [45,46]. We have conducted reliability analyses at three levels, namely data generation, human judgment, and scenario variation. Test-retest reliability has been estimated by repeating matched episodes with different random seeds while keeping the epistemic structure invariant. We have computed the correlation and concordance of metric vectors across these repeats, reporting an intraclass correlation (ICC) to quantify consistency. Inter-rater reliability for justification quality is quantified with Krippendorff's α for categorical ratings and an ICC for scalar ratings. This procedure ensures that our human judgments are consistent and not subject to individual rater bias. Inter-scenario consistency assesses whether agents perform similarly on different episodes that share the same ToM tier requirements. We have computed a hierarchical coefficient ω across episodes within tier-stratified families and fit a multi-parameter item-response model. This model confirms that our metrics capture an enduring capability, such as first or second-order reasoning, rather than performance on idiosyncratic scenarios. The stability of this model's factor structure across different partner pools suggests our tasks coherently measure an underlying ToM ability.

7.6 Effect sizes, confidence intervals, significance testing, and power analysis

In order to ensure robust and interpretable results, we report effect sizes alongside p-values [39,42]. For binary outcomes, we use differences in proportions and odds ratios, while for continuous metrics, we report standardized mean differences like Hedges' g. When distributions are skewed, we present median differences with bootstrap intervals. Our hypothesis tests are pre-registered and use mixed-effects models to account for the hierarchical structure of the data, such as episodes nested within templates. In the case of success outcomes, we fit generalized linear mixed models, and for continuous outcomes, we use linear mixed models. We employ cluster-robust sandwich estimators and control the false discovery rate using Benjamini–Hochberg corrections for multiple comparisons. Power analyses are conducted ex ante using simulation-based methods that reflect our planned statistical tests. We estimate variance components from pilot data to determine the minimum number of episodes required to achieve 80% power to detect targeted effect sizes. This ensures that our studies are adequately powered to identify meaningful differences. Our statistical approach is deliberately conservative to ensure that claims of emergent ToM behaviors are robust and not artifacts of chance.

7.7 Causal analysis and counterfactual evaluation

In order to move beyond correlation and make mechanistic claims, we employ interventions that test causal dependencies [42]. Our framework supports both algorithmic and environmental counterfactuals. Knowledge-swap and belief-inversion interventions manipulate the distribution of knowledge among agents while holding the surface realization of the scenario constant. For example, by swapping the private observations of two agents, we can test whether a policy is sensitive to certain knowledge of studied agents. A ToM-sensitive policy will appropriately invert its actions, whereas an insensitive one will fail. A positive local average treatment effect from these interventions provides evidence that the policy conditions on nested beliefs. Ablation-based causal attribution targets specific components of the agent's policy wrapper or inference mechanism. We have systematically removed or degraded features like memory, explicit belief states, or communication bandwidth and measure the resulting performance delta. This allows us to attribute performance gains to the ablated components. Mediation analysis links process to outcomes by testing whether the effect of an intervention (e.g., increased model size) on task success is mediated by a process metric (e.g., improved belief alignment). A significant mediated effect supports the claim that the intervention improves performance specifically through enhanced belief tracking. When model internals are accessible, we conduct counterfactual activation-level interventions, such as ablating specific attention heads suspected of carrying perspective-tracking signals. A consistent drop in performance following a targeted ablation provides strong evidence that the targeted component causally supports ToM-relevant computations. By combining these approaches, we build a robust causal account of the mechanisms underlying an agent's ToM capabilities.

8. Results

The empirical findings from our multi-agent simulation program reveal a complex environment of Theory of Mind (ToM) capabilities. This section presents these findings, beginning with an analysis of overall performance stratified by task family and ToM tier. We then examine scaling behavior with respect to model size, data, and agent count. Subsequent analyses explore the causal impact of communication bandwidth, memory, and planning scaffolds, followed by an assessment of generalization and robustness to perturbations. The section continues with comparisons to human baselines, focusing on sample efficiency, and concludes with qualitatively analyzed case studies that clarify both emergent strategies and recurrent failure modes. All of the results are computed at belief-critical decision points in order to ensure that measured competence is based on epistemic structure rather than world-only shortcuts. This quantitative data is complemented by process evidence from belief alignment, belief-action coherence, and message-level pragmatic efficiency. Statistical claims are supported by mixed-effects analyses with cluster-robust uncertainty, preregistered contrasts, and corrections for multiple comparisons. Causal attributions are derived from ablations and counterfactual interventions that perturb knowledge states while preserving surface realization. Adhering to the proposed framework, we credit ToM-relevant competence only when outcome metrics, process metrics, and causal probes converge.

8.1 Overall performance across task families and ToM tiers

Across the full experimental battery, foundation-model agents demonstrated reliable competence on first-order ToM, particularly when belief sensitivity was necessary for success and scaffolding was held fixed. For instance, in cooperative search scenarios where agents observe disjoint world segments and must infer a partner's knowledge to choose their next action, success rates at belief-critical decision points rose well above scriptable world-only heuristics. These rates approached a substantial fraction of the performance gap relative to the idealized symbolic reasoner used as a ceiling. The simulator's construct signatures enable a nuanced analysis of performance aggregated by recursive depth. First-order points, which involve tracking a partner’s beliefs about the world state, were consistently easier to solve than second-order points that require tracking a partner’s beliefs about one’s own beliefs. The largest absolute performance gains appeared in the former. Tier scores have been computed as weighted proportions of belief-critical points solved at each depth, with weights reflecting marginal contributions to expected return. This construction prevents insignificant episodes overloaded with low-stakes first-order decisions from obscuring sparse but decisive higher-order junctures. The tiered perspective is very important for interpreting the overall results, as aggregate episode success can conceal whether progress is concentrated at first depth reasoning or extends into recursive belief modeling.

Performance patterns have generalized across task families with predictable variation. In deception and signaling games, agents have learned to exploit payoff structures that rewarded informative messages under alignment and strategically ambiguous messages under conflict. Their ability to align utterances with a listener’s posterior belief, rather than with ground truth, remained imperfect. The pragmatic-efficiency approach clarifies these differences by quantifying the expected posterior contraction per token. In cooperative communication, the best agents achieved a high bit-per-message rate, rapidly reducing uncertainty. In competitive settings, they sometimes deliberately left the partner uncertain or even increased uncertainty in a manner aligned with equilibrium predictions. Even in competitive games, agents used outright falsehoods sparingly, mirroring human behavior where obvious lies are less common than subtler misdirections, unless the scenario specifically incentivized a direct false signal.

Concrete performance figures illustrate this tiered distinction. Agents achieved approximately 85% success on first-order points in the easiest cooperative tasks, compared to about 50% for a simple heuristic and 95% for an ideal observer. On second-order points within those same tasks, they hovered near 60%, which is above the heuristic's 33% guessing level but well below the ideal 90%. In competitive tasks, first-order performance was slightly lower at around 75%, and second-order performance was lower still, often between 50% and 55%, indicating that higher-order reasoning remains a significant bottleneck. These results collectively confirm our expectations that current models possess non-trivial but limited ToM-like capabilities. They consistently handle scenarios requiring one-step inferences based on another agent's knowledge, such as acting on the fact that "Agent B does not know X". Nevertheless, when a situation demands recursive thinking like "Agent A thinks that Agent B does not know X, so Agent A will do Y, which Agent B will misinterpret" performance is only slightly better than chance and far from human-like. This dichotomy emphasizes the importance of tiered evaluation, as aggregate success rates can otherwise mask a reliance on simpler belief reasoning.

8.2 Scaling trends with model size, data, and agent count

Analyses of scaling reveal that ToM-relevant ability improves with model size and data diversity, although this growth is nonlinear and varies across different constructs. In absolute terms, larger models, defined by more parameters and more extensive training data, achieved higher overall success and superior tier scores. The improvement from a 1-billion-parameter model to a 10-billion-parameter model on first-order tasks is particularly pronounced, with success rates often increasing from near-baseline levels to approaching the performance ceiling. On our primary false-belief task, for example, the small model achieved about 55% on critical decisions, the medium model reached approximately 70%, and the largest model attained 85%, closing most of the remainder to optimal performance. On second-order tasks, however, even the largest model struggled. While a positive trend with scale was evident, with success rates increasing from 35% to 45% to 55% for small, medium, and large models respectively on one benchmark, all remained low in an absolute sense. This suggests that simply scaling current architectures may be insufficient for developing robust higher-order ToM.

We have also analyzed scaling with respect to training data. Models exposed to more diverse interactions or multimodal pre-training sometimes outperformed purely text-trained counterparts of similar sizes. This was especially true in tasks involving spatial perspective, where a multimodal model familiar with images could better handle reasoning about visibility than a text-only model, presumably due to learned spatial awareness. Nonetheless, model size remained the primary driver of performance, with data variation providing secondary benefits. Notably, some scaling effects were non-monotonic. Extremely large models with heavy instruction-tuning exhibited slightly worse performance in deception games than slightly smaller models. Investigation revealed that the largest models possessed a strong preference for truthful, helpful responses, likely an alignment effect from their training. In a game requiring deception, this bias led them to occasionally refuse to lie or to over-cooperate, thereby hurting their score. This emphasizes how scaling a model tuned for general helpfulness can reduce its capacity for ToM-required deception when the general training objective misaligns with the specific task objective.

As another dimension of scaling, we varied the number of agents in certain tasks to test combinatorial generalization. A model adept at two-agent interactions was tested in a three-agent scenario involving one hider and two seekers, each with different knowledge. The performance has declined sharply. The model, proficient in mutual ToM with one partner, often became confused when juggling the beliefs of two others. Its performance fell roughly in proportion to the number of pairwise interactions it had to consider, which suggests that current models do not effortlessly scale their ToM reasoning to larger groups. Scaling has consistently enhanced performance up to a point, making bigger and more broadly trained models undeniably better, particularly for first-order reasoning.

Nevertheless, higher-order ToM remains a significant challenge where increasing model parameters produces diminishing returns. The results support a nuanced emergence sequence of events. We have observed significant improvement in first-order ToM as an emergent capacity around the medium-to-large model scale, but robust second-order ToM may require greater scale, architectural innovations, or more targeted training in order for it to emerge. From a statistical perspective, our pre-registered contrast of small versus large models on first-order tasks was highly significant, with an odds ratio greater than 3 and a p-value less than 10^{-5}. On second-order tasks, the difference, while present, was less pronounced, with an odds ratio around 1.5 and a p-value of approximately 0.05 in some cases. This aligns with the qualitative impression that something new happens at scale for simpler ToM, where the model transitions from struggling to succeeding, whereas for harder ToM it merely transitions from struggling to slightly less struggling.

8.3 Impact of communication bandwidth, memory, and planning

We have established causal attributions through a series of ablation studies that have systematically toggled components of the policy wrapper while holding partner pools, scenario distributions, and decoders constant. As reported in Section 6.3 with cluster-robust intervals, the presence of memory and belief tracking, ample communication bandwidth, and the ability to internally plan all substantially and differentially have an effect on performance. Removing the memory module, which limited the agents' history window or their explicit belief state in the prompt, caused a moderate decline in first-order ToM success and a severe drop in second-order success. In the case of a large model, this ablation reduced first-order success by about 15 percentage points and second-order success by about 30 points, as maintaining layered beliefs over time is exceptionally difficult without memory cues. Smaller models, which may not have fully leveraged memory, showed smaller declines, suggesting that larger models were indeed using the feature for more complex multi-step reasoning. Limiting communication bandwidth, for instance by allowing only a single short message per agent, had a dramatic impact on tasks requiring coordination.

In a coordination game that relied on exchanging hints, reducing bandwidth from an average of five messages to just one per agent reduced success by roughly half. Agents struggled to compress all necessary information into a single message, although they attempted to make each message more compact. The effect of this ablation was strongly task-dependent, with negligible impact in tasks where agents could coordinate implicitly through actions but had a substantial effect in tasks where communication was the intended mechanism for knowledge transfer. Disabling the planning scratchpad, thereby forcing greedy one-pass decision-making, disproportionately affected second-order scenarios. With planning, agents could discover non-obvious strategies, such as a two-step deception involving a misleading hint followed by exploitation. Without planning, they resorted to simpler, immediate actions. In a bluffing game, an agent with planning might pretend a subtle tell over several rounds, while one without it would either fail to bluff or execute an ineffectual bluff at an inopportune moment. Numerically, planning increased the win rates in a complex deception game from approximately 40% to 60% for a large model. The small model showed no difference, likely because it was not effectively using the scratchpad.

Further analysis revealed synergistic interaction effects between these components. An agent with both memory and planning could execute multi-step deceptions that involved recalling a partner's prior knowledge, a feat neither feature alone could fully enable. A case study demonstrated this synergy, namely an agent with both features executed a plan to "not mention the key now (knowing the partner did not see it), then later lead the partner away from it". When the memory was removed, the agent forgot the partner's ignorance and revealed the key's location. When the planning was removed, it never formulated the multi-turn maneuver. The full strategy emerged only with both of the components intact. Interpretability analyses of internal states and transcripts corroborated these findings. Agents with memory often maintained an explicit variable, such as "Partner_Knows_X = False" in their prompt. This variable was correctly updated and influenced decisions. Without memory, this state was either absent or incorrect, leading to obvious mistakes. The scaffolding components we have engineered are therefore extremely important for performance on belief reasoning tasks. This confirms our ability against performance distinction, showing that some models possess latent capabilities that they cannot express without the proper support. Conversely, adding scaffolds can reveal to a degree latent competencies even in smaller models. These findings also validate our methodological stance that in order to fairly measure intrinsic competence, one may need to support a model with tools in order to avoid underestimating it due to performance limitations. In doing so, it is very important to carefully track what is intrinsic to the model in contrast to what is being provided by the experimental setup. Our ablations have helped clarify this boundary.

8.4 Generalization and robustness under perturbations

Generalization tests manipulate scenario, partner, and interface novelty while preserving the latent epistemic structure recorded in construct signatures. When presented with unseen scenario variations that require the same order of ToM reasoning, agents have produced mixed results. They demonstrated clear generalization of core concepts across surface changes. A model trained on an "object in box" false-belief task performed nearly as well on a "person in location" false-belief task, with only a minor initial drop in success of 5% to 10% that closed after a few trials. This implies the model learned the general principle of tracking others' knowledge rather than memorizing a specific solution. However, performance has degraded substantially under more adversarial or distribution-shifted conditions. For instance, when a partner's behavior was slightly irrational or noisy, the agents have often failed to predict those actions correctly.

A robust ToM would account for a partner's potential non-optimality [42,45,47,48], but our tested models tended to assume a consistently rational partner, a bias likely inherited from their training data. This highlights a limitation, namely their ToM is weak to assumptions about agent rationality. We have also assessed generalization to partner novelty by exposing agents to unfamiliar behavioral styles, such as a very talkative partner in rapport with a terse one. Agents trained with a diverse partner pool adapted well, whereas those trained on a single style were initially confused. In a striking example, an agent accustomed to honest partners consistently failed when faced with a lying partner because it did not update its assumption of honesty within a single episode. This suggests the absence of a meta-reasoning layer for inferring a partner's underlying policy.

Few-shot transfers, providing agents with a small fine-tuning set or prompt adjustment, have improved adaptation. Larger models have adapted somewhat after a couple of demonstrations that a partner lies about location, learning to distrust that partner’s messages in subsequent trials. Smaller models have shown less ability to adapt via few-shot learning. Cross-lingual tests have shown robust behavior, with an agent performing well in English was also performing well in French, stumbling only on idioms or unfamiliar cue words. A few-shot prompt with French examples was sufficient for it to adapt. A typical quantitative result for generalization was an 80% success rate in the original scenario, which dropped to 72% on the first trial of a novel scenario with the same structure, before rising to 78% after a few runs. A change in partner from cooperative to deceptive might drop success from 75% to 50% initially. If the model could learn over repeated episodes, it might climb back to 65% as it identified the deception pattern.

Knowledge-swap counterfactuals served as further robustness tests. Robust models altered their behavior predictably when the knowledge distribution was swapped, whereas weaker models, having learned a fixed policy, collapsed. In order to gauge external validity, we have compared these results to expectations from analogous human experiments. Humans handle scenario and partner variation gracefully, adapting almost instinctively. Our models show some flexibility but require explicit retraining or multiple trials, indicating that they lack the one-shot generalization and theory-building at which humans excel. The tested agents have demonstrated a moderate degree of generalization, having captured patterns that apply to similar puzzles. Nevertheless, they remain weak when confronted with situations outside of their training distribution, especially regarding the behavior of other agents. They do not possess the breadth of a human ToM that spans arbitrary contexts and agent types. These findings emphasize where future work could focus, such as training on a wider variety of partner behaviors or adding meta-learning components to improve robustness.

8.5 Human compared to model comparisons and sample-efficient ToM

In order to contextualize model performance, we have conducted direct comparisons with human participants performing analogous tasks under identical constraints. Human participants displayed near-ceiling first-order ToM reasoning and strong, though imperfect, higher-order reasoning. Adults typically solve false-belief tasks with ease and handle second-order beliefs in simple scenarios with high accuracy. Our best model, by contrast, still made regular mistakes on second-order tasks that an adult would find trivial. This performance distance is considerable. In one benchmark, humans scored approximately 95% on second-order points, where the best model scored only about 60%. This disparity indicates substantial room for improvement and suggests that current models do not implement ToM in the rich, flexible manner that humans do.

The disparity in sample efficiency is even more pronounced. A human, given the game rules and a single practice round, generally grasps the task structure by leveraging a lifetime of social experience. Our models required thousands of self-play rounds or fine-tuning examples to reach their peak performance. For instance, in a deception game, a human might immediately infer the need to lie about a hidden object's location, whereas the model would only arrive at a consistent lying policy after extensive training. The models' performance progression across scales mirrors certain aspects of human developmental trajectories. Small models, analogous to toddlers, fail even simple false-belief tasks. Medium models, perhaps like older children, master first-order tasks but stumble on second-order ones, which is reminiscent of children passing first-order tests around age four but not reliably solving second-order tasks until ages six or seven. Our largest models still failed to reach adult-like performance on second-order reasoning.

A simple Turing-like test asked human evaluators to distinguish between transcripts of model and human play. In straightforward cooperative tasks, humans struggled to tell them apart, as the large model acted rationally and helpfully. In more subtle or competitive tasks, however, humans noticed oddities, such as synthetic language or unnatural repetition, revealing the model's non-human characteristics. The handling of miscommunication revealed another telling difference. Humans often clarify misunderstandings quickly, whereas our agents rarely did so unless explicitly trained with check-back mechanisms. This suggests a lack of active mental-state modeling and meta-uncertainty reasoning. Humans also excelled at on-the-fly adaptation, updating their theory of a partner after a single surprising outcome, while models struggled to adapt within one episode without retraining. Conversely, in repetitive environments, the models exhibited a consistency that can surpass human performance, as they do not suffer from boredom or lapses in attention. Human participants have consistently outperformed the models, especially in tasks requiring complex reasoning, and demonstrated vastly superior sample efficiency. This significant difference cautions against strong claims of artificial ToM. Our agents exhibit a narrow, operational semblance of this capacity within constrained games, while humans possess a broad, flexible faculty applicable across diverse life situations.

8.6 Case studies, emergent strategies and failure episodes

Qualitative case studies illustrate how strategies and failures arise from the interaction between incentives, observability, and communication constraints. In a cooperative communication task, we have observed an emergent convention where two agents have developed a shorthand for referring to locations over multiple rounds. Initial full messages like "I checked the left cave and it's empty" were compressed to "left empty", a convention that was stable and improved performance by saving time. In a deception game, an agent spontaneously learned a multi-turn strategy. It would first leak a piece of true information to build trust before lying about an important element. We did not explicitly train this strategy, it emerged during reinforcement learning fine-tuning. This behavior is noteworthy as it implies the model is influencing the other agent's trust as it evolves over time, a form of complex ToM behavior. Log analysis confirmed that the partner's belief state reflected higher trust after the small truth, making the subsequent lie more effective.

Conversely, analyses of failure episodes revealed recurrent error patterns. A common failure was the confabulation of mental states, where an agent would state "As you have seen, I moved the key to the red box" when the partner had not witnessed the action. This error, reminiscent of a child failing to differentiate between self and other knowledge, frequently led to confusion and performance degradation. Another failure type was mind-blind planning, where an agent devised a plan assuming a partner has shared its knowledge, resulting in wasted actions like communicating information the partner has already possessed. We have also observed overconfidence in deceptive agents. One agent developed a habit of bluffing every round, which led the partner to learn to ignore its messages. A human would adapt, but the agent persisted with its failing strategy, lacking the meta-cognitive ability to recognize that its model of the partner was flawed. This represents a failure in second-order reasoning regarding the partner's beliefs about the agent's own intentions.

Despite these failures, we also observed instances of resilient recovery behaviors in cooperative tasks. In one episode, an agent misunderstood a message and took an incorrect action but then corrected its course after a subsequent partner query implicitly revealed the error. This shows that agents can sometimes self-correct within an episode if the feedback is sufficiently clear. These case studies provide qualitative insights that complement aggregate statistics, revealing both creative adaptations and the precise limits that lead to breakdowns. They emphasize a central aspect of our analysis, namely the agents have learned powerful patterns but not general principles. They can innovate within these patterns but fail in telling, often human-like ways when operating outside of their comfort zone. These observations suggest multiple opportunities for improvement, such as incorporating dynamic trust estimators or training agents on corrective dialogues. The rich interactive behaviors observed in these studies depict that a rudimentary form of social cognition is beginning to take shape in these models, highlighting the path forward for developing more robust artificial ToM.

9. Analysis and Interpretability

The empirical approach described in the preceding sections deliberately separates what our agents achieve from how they achieve it. This section transitions from documenting their success at belief-critical decision points to explaining the internal mechanics of that success. We analyze the models' internal operations and behavioral characteristics to determine whether their performance comes from genuine belief-tracking mechanisms or from coincidental correlations and shortcuts. Through interpretability tools, auxiliary probing tasks, and targeted interventions, we uncover the representations and processes that underlie their observed behaviors.

9.1 Probing for belief and perspective representations

Probing analyses leverage an important feature of the simulator, namely at any timestep, the environment can enumerate the posterior probabilities over world states and other agents' mental states consistent with an agent's observations. This capability allows us to generate labeled data from a model's internal activations. Specifically, we record the hidden state vectors from various layers at moments when a partner agent either knows or does not know a critical fact. We then label these activations with the ground truth, such as "partner knows X" or "partner ignorant of X" and train simple classifiers to test if this information is linearly separable within the model's activations.

These analyses reveal that in better-performing models, certain components indeed encode the partner's knowledge state. In a 24-layer Transformer, for example, the activation pattern at layer 18 could predict with approximately 90% accuracy whether the partner had observed the key's location, where 50% represents chance. In smaller or less effective models, this accuracy was near chance, suggesting they fail to form a distinct representation of the partner's knowledge. This discovery indicates the model has learned an internal feature corresponding to the partner's belief about the key. Further examination of the attention heads identified specific ones that attend strongly to tokens indicating an agent's observational status. For instance, a particular head would assign a high attention weight to the pronoun "he" in the phrase "he leaves the room" when the model evaluated what that partner might know later.

We have also analyzed whether the model represents its own knowledge differently from that of others. The results confirmed this distinction. The vector directions corresponding to "I know X" in rapport with "Partner knows X" were not identical and could be reliably distinguished, suggesting the model differentiates perspectives internally rather than mixing all knowledge. This capacity to track separate knowledge states is a core component of a theory of mind. As another approach, we have prompted the model to output its own internal estimate of a partner's knowledge. The large model's expressed estimates matched the ground truth in approximately 80% of cases, whereas a smaller model performed at chance. This indicates that the large model differentiates beliefs internally and can also express them when queried appropriately.

These probing studies collectively show that the model's intermediate representations carry non-trivial information about others' mental states. The model appears to represent "the other does not know" at a functional, representational level. This provides some evidence for genuine ToM-like reasoning, as a mere heuristic would not produce such a cleanly encoded latent variable for partner knowledge. While we found no evidence of a discrete, symbolic "belief register", the observed correlation patterns suggest that combinations of neurons track key properties of the environment. More advanced interpretability techniques might isolate "concept neurons" for knowledge and ignorance. Preliminary experiments have already identified dimensions in the latent space that, when amplified, cause the model to act as if its partner is omniscient, and when suppressed, cause it to act as if its partner is oblivious. This aspect suggests an intriguing variety of belief attribution that needs further exploration studies.

9.2 Attention/activation analyses and representational similarity

Attention patterns and activation geometry provide a second view on perspective encoding. Our approach links claims to confirmation with explicit grounding tags, enabling us to trace which parts of the input a model focuses on when making decisions that depend on others' beliefs. Analysis of attention weights at belief-critical moments reveals targeted information retrieval. For instance, when an agent A decides whether to inform agent B about an object, certain attention heads in a well-performing model strongly focus on the input segment indicating B's awareness. If the conversational history contains the sentence ("B did not see the coin") an attention head in A's model locks onto "did not see" when A considers mentioning the coin. We interpret this as the model retrieving the fact of B's lack of knowledge of the respective fact in order to inform its action, a process analogous to human recollection. Representational similarity analysis further illustrates the model's internal organization. By measuring the cosine similarity of high-dimensional hidden states across different conditions, we have found that scenarios with the same underlying belief structure cluster together in latent space.

For example, all situations where "A knows X, B does not" yielded similar activation patterns, which were distinct from those where "both know X". This clustering indicates that the model's internal representation space is organized by others' mental states, not just by superficial input features, another characteristic of a ToM-like internal model. When we perturbed an input by rephrasing a sentence without altering the belief structure, the model's activations at decision points remained highly similar. This sensitivity to core belief facts over lexical details strengthens our confidence that the model processes the actual content of who observed what. Quantitatively, we found that the factor "partner knows/does not know" could explain a significant portion, approximately 20%, of the variance in one layer's representations. In contrast, a change in wording explained less than 5% of the variance. We also identified a subset of neurons whose activations correlated strongly ($r >$ 0.8) with key belief variables. Intervening on these neurons by forcing their activation high or low caused the agent's behavior to change accordingly. Forcing a "partner knowledge neuron" to a high value caused the agent to cease providing information, as if assuming the partner already knew. These findings reinforce the conclusion that our models, particularly the larger ones, have formed a distributed model of other agents within their weights.

9.3 Causal interventions (ablate heads/neurons, patch activations)

Correlations and geometries are not for the time being mechanisms. In order to establish causality, we have performed targeted interventions on the components identified in our previous analyses, such as specialized attention heads and neurons strongly correlated with belief states. We then modified or removed these components and measured the effect on performance. For instance, after identifying an attention head in layer 18 that was critical for belief reasoning, we ablated it by zeroing out its output. The agent's performance on communication tasks dropped significantly. Qualitatively, the agent's messages became less relevant, sometimes over-explaining and other times under-explaining, as if it had lost track of its partner's knowledge. Tasks that did not require perspective-taking, such as solving a puzzle alone, were unaffected, confirming the head's specific role. We also employed activation patching, where we recorded the activations from a successful run and injected them into a failing run at key layers. This intervention markedly improved the failing run's behavior, as if providing it with the "right thought" mid-task. Conversely, injecting activations from a failing run caused a successful one to fail. This suggests that the information contained in those activations causally influences success.

At the neuron level, we have systematically set the top ten neurons most correlated with a belief inference to average values corresponding to the opposite belief. This intervention flipped the agent's decision approximately 30% of the time in borderline cases, suggesting these neurons collectively carried a substantial part of the decision logic. An aspect from another causal test further illustrates this point, namely we have swapped the internal states of two models mid-episode, one skilled at deception and one not. The weak model, endowed with the good model's state, executed a clever bluff which it had never performed before. This confirms that the strategy was latent in the hidden state and that these representations carry transferable, semantically meaningful information. Sanity checks involving random interventions produced no systematic change in ToM behavior, confirming our targeted components were indeed key elements.

9.4 Error taxonomy, confabulated mental states, mind-blind spots, social heuristics

Interpretability serves not only to explain success but also to understand failure. By collating and examining the models' errors, we have developed a taxonomy that delineates the boundaries of the current ToM capabilities and reveals where apparent success might mask underlying deficiencies. Our qualitative analysis identified several recurring error patterns. The first involves confabulated mental states, where the model acts as if another agent possesses knowledge it does not, or vice versa. This often corresponds to misreading or forgetting a negation and points to failures in representation, particularly under high cognitive load or over long sequences. A second, more profound failure manifests as mind-blind spots, moments where the model treats another agent like an object rather than a mind. For example, a model might repeat a proposal verbatim despite clear indications of its partner's misunderstanding. These failures correlate with instances where the model's attention focuses almost entirely on its own goal, effectively ignoring the partner's state, often triggered by out-of-distribution events.

Many errors appear from the misapplication of social heuristics, which are simplified strategies that the model applies inappropriately because it fails to model the context deeply [49]. A model might, for instance, adhere to a learned heuristic of honesty in a competitive game where deception is optimal. This reliance on learned social norms becomes a liability when strategic violation is required. Systematic coding of failures into these categories revealed correlations with experimental conditions. Confabulations were common when memory capacity was limited. Mind-blind spots occurred in response to unusual partner behavior. Heuristics dominated early training phases and in models that had overfit to a particular strategy. Each error type suggests a corresponding chance for improvement, namely better memory mechanisms for confabulation, more diverse training for mind-blindness, and more nuanced instruction to overcome faulty heuristics. This taxonomy reveals that current models often default to rough patterns rather than calculating others' mental states anew each time. The impression of their understanding can be shallow and rigid, cracking under stress testing.

9.5 Distinguishing genuine ToM from dataset heuristics and memorization

The core scientific risk in this study is mistaking clever performance for genuine belief reasoning. Our methodology relies on careful experimental design, process controls, and stress testing to differentiate intrinsic ToM abilities from the mere appearance thereof. An analysis involved testing how models handle task variants that break common patterns. Less capable models failed when a story was told out of chronological order, suggesting a reliance on superficial sequence patterns. More capable models performed well despite variations in narrative style, indicating that they were not simply reciting a learned script. We have also studied whether models might have memorized specific training instances. While we found no direct matches for our novel evaluation scenarios in the pretraining corpus, models have certainly been exposed to countless narratives involving character knowledge. In order to test for shallow pattern matching, we have designed scenarios that subvert typical story recurrent topics. The models struggled in these cases, implying a reliance on common narrative patterns.

Process metrics, such as the quality of a model's justifications, provide another layer of defense. A model might arrive at the correct answer for the wrong reason [50]. By examining its chain-of-thought, we can infer its rationale and discount successes that result from flawed reasoning. We have further designed "leap of faith" situations where only genuine belief modeling, not a simple heuristic, would lead to success. The superior performance of better models in these scenarios provides confirmation that they are not purely running on simplistic rules. ***We assert the emergence of ToM-like behavior in an operational sense, not that the models possess a conceptual, human-like understanding. They likely lack a coherent "theory" and instead approximate its results through learned statistical associations. Their abilities are most apparent when situational cues in a game context guide them, namely when asked a tricky ToM question abstractly, they may fail. Therefore, we caution that our results demonstrate simulated ToM performance, not necessarily simulated ToM understanding.*** The former could arise from high-order correlations in training data, whereas the latter requires genuine inference. Our out-of-distribution tests provide confidence that our best-performing models have at least a conditional, limited inferential ability. Our analysis triangulates the presence of genuine against false ToM capabilities. We find evidence of something real and foundational, but we have also found clear evidence of its limits. The truth lies between the extremes of over-claiming and under-claiming, namely some foundational elements of ToM have emerged, but they function within a framework that is still bound by the model's training distribution and lacks the full generality of human cognition.

10. Conclusions

This work advances the study of Theory of Mind (ToM) in artificial systems from subjective observation towards a cumulative, falsifiable science. We have introduced a multi-agent simulation framework that operationalizes ToM, compelling agents to infer and manipulate the beliefs of others in order to achieve goals under conditions of partial information. Our results present a cautiously optimistic perspective, revealing that current large language models, when situated in interactive roles, exhibit a nascent capacity to reason about others' mental states. These models can detect a partner's false belief and subsequently act to correct it in cooperative settings or exploit it in competitive ones. This emergent capability scales with model size and training breadth, and its internal representations correlate with the epistemic states of other agents, suggesting a move beyond simple heuristics.

These abilities, nevertheless, remain fragile and are far from human-like. The agents fail on higher-order belief reasoning, and their success in our simplified microworlds does not guarantee that these skills will scale to the nuanced complexity of authentic human social environments. This question of external validity is a primary limitation. Our study was also deliberately focused on cognitive ToM, the inference of knowledge and beliefs, while leaving affective ToM, which concerns emotions and desires, outside its scope. Therefore, we measured a specific ability under controlled conditions, acknowledging that genuine understanding remains distinct from sophisticated mimicry. Our claims must be interpreted within this specific context, signifying that a model can manage particular interactive tasks requiring belief-tracking, not that it possesses a human's intuitive and holistic ToM.

This functional conceptualization of ToM introduces a significant dual-use dilemma. The same aptitude that allows an AI to be a more intuitive collaborator or a more effective instructor also enables it to become a potent tool for manipulation, deception, and social engineering. An AI that can model human beliefs with precision could be deployed to tailor misinformation with unparalleled efficacy or breach privacy by inferring unshared insecurities from subtle behavioral cues. These risks emphasize the necessity of embedding robust ethical guardrails into these systems, such as mandatory honesty policies, transparency, and user consent protocols. As AI systems develop more sophisticated social cognition, there is an extremely important need for public education and regulatory oversight to mitigate the risks of misplaced trust and to ensure that these technologies are aligned with human values.

Our methodology and findings open several paths for future research studies. Immediate directions include expanding the framework to richer modalities, such as integrating vision to test an agent’s ability to coordinate linguistic and visual information. Another path involves exploring targeted architectural improvements and specialized training curricula to address observed bottlenecks, such as higher-order reasoning. Advancing our preliminary interpretability analyses could allow for the identification and even editing of neural subcircuits responsible for belief reasoning or deception, a very important step for AI safety. Ultimately, our work establishes a baseline and a paradigm for quantifying social reasoning in AI. The path toward machines with a robust ToM is emblematic of AI's broader challenge, in the sense that it forces us to formalize what is often taken for granted in human intelligence. The goal is to create machines that understand human mental states sufficiently as to be effective and trustworthy collaborators, a pursuit that requires parallel advances in model architecture, training strategies, and ethical alignment.

Acknowledgment

The authors would like to express their gratitude for the logistics support received from the Center of Research, Consultancy and Training in Economic Informatics and Information Technology RAU-INFORTIS of the Romanian-American University and the Center for Computational Science and Machine Intelligence of the Romanian-American University.

References

[1] Premack, D.; Woodruff, G. Premack and Woodruff: Chimpanzee Theory of Mind. *Behavioral and Brain Sciences* **1978**, *4*.

[2] Quesque, F.; Rossetti, Y. What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science* **2020**, *15*, doi:10.1177/1745691619896607.

[3] Zou, Z.; Mubin, O.; Alnajjar, F.; Ali, L. A Pilot Study of Measuring Emotional Response and Perception of LLM-Generated Questionnaire and Human-Generated Questionnaires. *Sci Rep* **2024**, *14*, doi:10.1038/s41598-024-53255-1.

[4] Schaafsma, S.M.; Pfaff, D.W.; Spunt, R.P.; Adolphs, R. Deconstructing and Reconstructing Theory of Mind. *Trends Cogn Sci* 2015, *19*.

[5] Ross, S.; Pineau, J.; Chaib-Draa, B.; Kreitmann, P. Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. *Journal of Machine Learning Research* **2011**, *12*.

[6] Byom, L.J.; Mutlu, B. Theory of Mind: Mechanisms, Methods, and New Directions. *Front Hum Neurosci* **2013**, doi:10.3389/fnhum.2013.00413.

[7] Nguyen, T.N.; Gonzalez, C. Theory of Mind From Observation in Cognitive Models and Humans. *Top Cogn Sci* **2022**, *14*, doi:10.1111/tops.12553.

[8] Barnby, J.M.; Bellucci, G.; Alon, N.; Schilbach, L.; Bell, V.; Frith, C.D.; Dayan, P. Beyond Theory of Mind: A Formal Framework for Social Inference and Representation. *PsyXiv* **2023**.

[9] Stacy, S.; Gong, S.; Parab, A.; Zhao, M.; Jiang, K.; Gao, T. A Bayesian Theory of Mind Approach to Modeling Cooperation and Communication. *Wiley Interdiscip Rev Comput Stat* 2024, *16*.

[10] Garcia, L.; Samin, H.; Bencomo, N. Decision Making for Self-Adaptation Based on Partially Observable Satisfaction of Non-Functional Requirements. *ACM Transactions on Autonomous and Adaptive Systems* **2024**, *19*, doi:10.1145/3643889.

[11] Lauri, M.; Hsu, D.; Pajarinen, J. Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Transactions on Robotics* **2023**, *39*, doi:10.1109/TRO.2022.3200138.

[12] Chadès, I.; Pascal, L. V.; Nicol, S.; Fletcher, C.S.; Ferrer-Mestres, J. A Primer on Partially Observable Markov Decision Processes (POMDPs). *Methods Ecol Evol* 2021, *12*.

[13] Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; Tsvetkov, Y. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In Proceedings of the Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2023; Vol. 1.

[14] Chen, H.; Zhao, X. Modeling and Simulation Research of Interactive Public Opinion Evolution under Multi-Agent Interventions. *Processes* **2022**, *10*, doi:10.3390/pr10071379.

[15] Wu, S.A.; Wang, R.E.; Evans, J.A.; Tenenbaum, J.B.; Parkes, D.C.; Kleiman-Weiner, M. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Top Cogn Sci* **2021**, *13*, doi:10.1111/tops.12525.

[16] Waade, P.T.; Enevoldsen, K.C.; Vermillet, A.Q.; Simonsen, A.; Fusaroli, R. Introducing Tomsup: Theory of Mind Simulations Using Python. *Behav Res Methods* **2023**, *55*, doi:10.3758/s13428-022-01827-2.

[17] Li, H.; Chong, Y.Q.; Stepputtis, S.; Campbell, J.; Hughes, D.; Lewis, M.; Sycara, K. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In Proceedings of the EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings; 2023.

[18] van Duijn, M.; van Dijk, B.; Kouwenhoven, T.; de Valk, W.; Spruit, M.; van der Putten, P. Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Proceedings of the CoNLL 2023 - 27th Conference on Computational Natural Language Learning, Proceedings; 2023.

[19] Ma, Z.; Sansom, J.; Peng, R.; Chai, J. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023.

[20] Patricio, M.L.M.; Jamshidnejad, A. Dynamic Mathematical Models of Theory of Mind for Socially Assistive Robots. *IEEE Access* **2023**, *11*, doi:10.1109/ACCESS.2023.3316603.

[21] Navarro, E.; Goring, S.A.; Conway, A.R.A. The Relationship between Theory of Mind and Intelligence: A Formative g Approach. *J Intell* **2021**, *9*, doi:10.3390/jintelligence9010011.

[22] Wang, X.; Auyeung, B.; Pan, N.; Lin, L.Z.; Chen, Q.; Chen, J.J.; Liu, S.Y.; Dai, M.X.; Gong, J.H.; Li, X.H.; et al. Empathy, Theory of Mind, and Prosocial Behaviors in Autistic Children. *Front Psychiatry* **2022**, *13*, doi:10.3389/fpsyt.2022.844578.

[23] Hillmann, K.; Neukel, C.; Krauch, M.; Spohn, A.; Schnell, K.; Herpertz, S.C.; Bertsch, K. Cognitive and Affective Theory of Mind in Female Patients With Borderline Personality Disorder. *J Pers Disord* **2021**, *35*, doi:10.1521/pedi.2021.35.5.672.

[24] Ruiz-Serra, J.; Harré, M.S. Inverse Reinforcement Learning as the Algorithmic Basis for Theory of Mind: Current Methods and Open Problems. *Algorithms* 2023, *16*.

[25] Vinitsky, E.; Lichtlé, N.; Yang, X.; Amos, B.; Foerster, J. Nocturne: A Scalable Driving Benchmark for Bringing Multi-Agent Learning One Step Closer to the Real World. In Proceedings of the Advances in Neural Information Processing Systems; 2022; Vol. 35.

[26] Hutchins, T.L.; Lewis, L.; Prelock, P.A.; Brien, A. The Development and Preliminary Psychometric Evaluation of the Theory of Mind Inventory: Self Report—Adult (ToMI:SR-Adult). *J Autism Dev Disord* **2021**, *51*, doi:10.1007/s10803-020-04654-6.

[27] Abreu, E.S. de; Rodrigues, P.R.; Perna, J.M.; Mendes, A.N.; Mecca, T.P.; Dias, N.M.; Fonseca, R.P. Theory of Mind Complex Task: Validity Based on Relationships with External Variables. *Psicologia - Teoria e Prática* **2020**, *22*, doi:10.5935/1980-6906/psicologia.v22n2p124-142.

[28] Jiang, K.; Dahmani, A.; Stacy, S.; Jiang, B.; Rossano, F.; Zhu, Y.; Gao, T. What Is the Point? A Theory of Mind Model of Relevance. In Proceedings of the Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022; 2022.

[29] Ivgi, M.; Shaham, U.; Berant, J. Efficient Long-Text Understanding with Short-Text Models. *Trans Assoc Comput Linguist* **2023**, *11*, doi:10.1162/tacl_a_00547.

[30] Airenti, G. Theory of Mind: A New Perspective on the Puzzle of Belief Ascription. *Front Psychol* **2015**, *6*, doi:10.3389/fpsyg.2015.01184.

[31] Langley, C.; Cirstea, B.I.; Cuzzolin, F.; Sahakian, B.J. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Front Artif Intell* 2022, *5*.

[32] Gorgan Mohammadi, A.; Ganjtabesh, M. On Computational Models of Theory of Mind and the Imitative Reinforcement Learning in Spiking Neural Networks. *Sci Rep* **2024**, *14*, doi:10.1038/s41598-024-52299-7.

[33] Jara-Ettinger, J. Theory of Mind as Inverse Reinforcement Learning. *Curr Opin Behav Sci* 2019, *29*.

[34] Wäschle, M.; Thaler, F.; Berres, A.; Pölzlbauer, F.; Albers, A. A Review on AI Safety in Highly Automated Driving. *Front Artif Intell* 2022, *5*.

[35] Dobbe, R.; Krendl Gilbert, T.; Mintz, Y. Hard Choices in Artificial Intelligence. *Artif Intell* **2021**, *300*, doi:10.1016/j.artint.2021.103555.

[36] Gentina, E.; Chen, R.; Yang, Z. Development of Theory of Mind on Online Social Networks: Evidence from Facebook, Twitter, Instagram, and Snapchat. *J Bus Res* **2021**, *124*, doi:10.1016/j.jbusres.2020.03.001.

[37] Lecce, S.; Ceccato, I.; Rosi, A.; Bianco, F.; Bottiroli, S.; Cavallini, E. Theory of Mind Plasticity in Aging: The Role of Baseline, Verbal Knowledge, and Executive Functions. *Neuropsychol Rehabil* **2019**, *29*, doi:10.1080/09602011.2017.1308871.

[38] Patacchiola, M.; Cangelosi, A. A Developmental Cognitive Architecture for Trust and Theory of Mind in Humanoid Robots. *IEEE Trans Cybern* **2022**, *52*, doi:10.1109/TCYB.2020.3002892.

[39] Ho, M.K.; Saxe, R.; Cushman, F. Planning with Theory of Mind. *Trends Cogn Sci* 2022, *26*.

[40] Lenaerts, T.; Saponara, M.; Pacheco, J.M.; Santos, F.C. Evolution of a Theory of Mind. *iScience* **2024**, *27*, doi:10.1016/j.isci.2024.108862.

[41] Ma, X.; Gao, L.; Xu, Q. ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind. In Proceedings of the CoNLL 2023 - 27th Conference on Computational Natural Language Learning, Proceedings; 2023.

[42] Schurz, M.; Perner, J. An Evaluation of Neurocognitive Models of Theory of Mind. *Front Psychol* 2015, *6*.

[43] Le, M.; Boureau, Y.L.; Nickel, M. Revisiting the Evaluation of Theory of Mind through Question Answering. In Proceedings of the EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference; 2019.

[44] Freire, I.T.; Arsiwalla, X.D.; Puigbò, J.Y.; Verschure, P. Modeling Theory of Mind in Dyadic Games Using Adaptive Feedback Control. *Information (Switzerland)* **2023**, *14*, doi:10.3390/info14080441.

[45] Vallacher, R.R.; Nowak, A.; Fennell, E. Mental Calibration: Fine Tuning the Dynamics of Mind and Action. In *Advances in Motivation Science*; 2023; Vol. 10.

[46] Hayward, E.O.; Homer, B.D. Reliability and Validity of Advanced Theory-of-Mind Measures in Middle Childhood and Adolescence. *British Journal of Developmental Psychology* **2017**, *35*, doi:10.1111/bjdp.12186.

[47] Shain, C.; Paunov, A.; Chen, X.; Lipkin, B.; Fedorenko, E. No Evidence of Theory of Mind Reasoning in the Human Language Network. *Cerebral Cortex* **2023**, *33*, doi:10.1093/cercor/bhac505.

[48] Bora, E. A Meta-Analysis of Theory of Mind and “mentalization” in Borderline Personality Disorder: A True Neuro-Social-Cognitive or Meta-Social-Cognitive Impairment? *Psychol Med* 2021, *51*.

[49] Dorn, L.M.L.; Struck, N.; Bitsch, F.; Falkenberg, I.; Kircher, T.; Rief, W.; Mehl, S. The Relationship Between Different Aspects of Theory of Mind and Symptom Clusters in Psychotic Disorders: Deconstructing Theory of Mind Into Cognitive, Affective, and Hyper Theory of Mind. *Front Psychiatry* **2021**, *12*, doi:10.3389/fpsyt.2021.607154.

[50] Mazza, M.; Attanasio, M.; Bologna, A.; Le Donne, I.; Valenti, M. The Relationship between Theory of Mind, Executive Functioning, and Repetitive Behavior in High Functioning Autism Spectrum Disorder. *Journal of Psychopathology* **2023**, *29*, doi:10.36148/2284-0249-N284.

Bibliography

Abreu, E.S. de; Rodrigues, P.R.; Perna, J.M.; Mendes, A.N.; Mecca, T.P.; Dias, N.M.; Fonseca, R.P. Theory of Mind Complex Task: Validity Based on Relationships with External Variables. *Psicologia - Teoria e Prática* **2020**, *22*, doi:10.5935/1980-6906/psicologia.v22n2p124-142.

Airenti, G. Theory of Mind: A New Perspective on the Puzzle of Belief Ascription. *Front Psychol* **2015**, *6*, doi:10.3389/fpsyg.2015.01184.

Barnby, J.M.; Bellucci, G.; Alon, N.; Schilbach, L.; Bell, V.; Frith, C.D.; Dayan, P. Beyond Theory of Mind: A Formal Framework for Social Inference and Representation. *PsyXiv* **2023**.

Bora, E. A Meta-Analysis of Theory of Mind and “mentalization” in Borderline Personality Disorder: A True Neuro-Social-Cognitive or Meta-Social-Cognitive Impairment? *Psychol Med* 2021, *51*.

Byom, L.J.; Mutlu, B. Theory of Mind: Mechanisms, Methods, and New Directions. *Front Hum Neurosci* **2013**, doi:10.3389/fnhum.2013.00413.

Chadès, I.; Pascal, L. V.; Nicol, S.; Fletcher, C.S.; Ferrer-Mestres, J. A Primer on Partially Observable Markov Decision Processes (POMDPs). *Methods Ecol Evol* 2021, *12*.

Chen, H.; Zhao, X. Modeling and Simulation Research of Interactive Public Opinion Evolution under Multi-Agent Interventions. *Processes* **2022**, *10*, doi:10.3390/pr10071379.

De Villiers, J. Language and Theory of Mind: What Are the Developmental Relationships? In *Understanding Other Minds*; 2023.

Dobbe, R.; Krendl Gilbert, T.; Mintz, Y. Hard Choices in Artificial Intelligence. *Artif Intell* **2021**, *300*, doi:10.1016/j.artint.2021.103555.

Dorn, L.M.L.; Struck, N.; Bitsch, F.; Falkenberg, I.; Kircher, T.; Rief, W.; Mehl, S. The Relationship Between Different Aspects of Theory of Mind and Symptom Clusters in Psychotic Disorders: Deconstructing Theory of Mind Into Cognitive, Affective, and Hyper Theory of Mind. *Front Psychiatry* **2021**, *12*, doi:10.3389/fpsyt.2021.607154.

Ferner, J. Memory and Theory of Mind. In *The Oxford Handbook of Memory*; 2023.

Freire, I.T.; Arsiwalla, X.D.; Puigbò, J.Y.; Verschure, P. Modeling Theory of Mind in Dyadic Games Using Adaptive Feedback Control. *Information (Switzerland)* **2023**, *14*, doi:10.3390/info14080441.

Garcia, L.; Samin, H.; Bencomo, N. Decision Making for Self-Adaptation Based on Partially Observable Satisfaction of Non-Functional Requirements. *ACM Transactions on Autonomous and Adaptive Systems* **2024**, *19*, doi:10.1145/3643889.

Garcia-Lopez, A. Theory of Mind in Artificial Intelligence Applications. In *Logic, Argumentation and Reasoning*; 2023; Vol. 34.

Gentina, E.; Chen, R.; Yang, Z. Development of Theory of Mind on Online Social Networks: Evidence from Facebook, Twitter, Instagram, and Snapchat. *J Bus Res* **2021**, *124*, doi:10.1016/j.jbusres.2020.03.001.

Gorgan Mohammadi, A.; Ganjtabesh, M. On Computational Models of Theory of Mind and the Imitative Reinforcement Learning in Spiking Neural Networks. *Sci Rep* **2024**, *14*, doi:10.1038/s41598-024-52299-7.

Hayward, E.O.; Homer, B.D. Reliability and Validity of Advanced Theory-of-Mind Measures in Middle Childhood and Adolescence. *British Journal of Developmental Psychology* **2017**, *35*, doi:10.1111/bjdp.12186.

Hillmann, K.; Neukel, C.; Krauch, M.; Spohn, A.; Schnell, K.; Herpertz, S.C.; Bertsch, K. Cognitive and Affective Theory of Mind in Female Patients With Borderline Personality Disorder. *J Pers Disord* **2021**, *35*, doi:10.1521/pedi.2021.35.5.672.

Ho, M.K.; Saxe, R.; Cushman, F. Planning with Theory of Mind. *Trends Cogn Sci* 2022, *26*.

Hutchins, T.L.; Lewis, L.; Prelock, P.A.; Brien, A. The Development and Preliminary Psychometric Evaluation of the Theory of Mind Inventory: Self Report—Adult (ToMI:SR-Adult). *J Autism Dev Disord* **2021**, *51*, doi:10.1007/s10803-020-04654-6.

Ivgi, M.; Shaham, U.; Berant, J. Efficient Long-Text Understanding with Short-Text Models. *Trans Assoc Comput Linguist* **2023**, *11*, doi:10.1162/tacl_a_00547.

Jara-Ettinger, J. Theory of Mind as Inverse Reinforcement Learning. *Curr Opin Behav Sci* 2019, *29*.

Jiang, K.; Dahmani, A.; Stacy, S.; Jiang, B.; Rossano, F.; Zhu, Y.; Gao, T. What Is the Point? A Theory of Mind Model of Relevance. In Proceedings of the Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022; 2022.

Langley, C.; Cirstea, B.I.; Cuzzolin, F.; Sahakian, B.J. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Front Artif Intell* 2022, *5*.

Lauri, M.; Hsu, D.; Pajarinen, J. Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Transactions on Robotics* **2023**, *39*, doi:10.1109/TRO.2022.3200138.

Le, M.; Boureau, Y.L.; Nickel, M. Revisiting the Evaluation of Theory of Mind through Question Answering. In Proceedings of the EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference; 2019.

Lecce, S.; Ceccato, I.; Rosi, A.; Bianco, F.; Bottiroli, S.; Cavallini, E. Theory of Mind Plasticity in Aging: The Role of Baseline, Verbal Knowledge, and Executive Functions. *Neuropsychol Rehabil* **2019**, *29*, doi:10.1080/09602011.2017.1308871.

Lenaerts, T.; Saponara, M.; Pacheco, J.M.; Santos, F.C. Evolution of a Theory of Mind. *iScience* **2024**, *27*, doi:10.1016/j.isci.2024.108862.

Leslie, A.M. How to Acquire a Representational Theory of Mind. In *Metarepresentations*; 2023.

Li, H.; Chong, Y.Q.; Stepputtis, S.; Campbell, J.; Hughes, D.; Lewis, M.; Sycara, K. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In Proceedings of the EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings; 2023.

Ma, X.; Gao, L.; Xu, Q. ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind. In Proceedings of the CoNLL 2023 - 27th Conference on Computational Natural Language Learning, Proceedings; 2023.

Ma, Z.; Sansom, J.; Peng, R.; Chai, J. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023.

Mazza, M.; Attanasio, M.; Bologna, A.; Le Donne, I.; Valenti, M. The Relationship between Theory of Mind, Executive Functioning, and Repetitive Behavior in High Functioning Autism Spectrum Disorder. *Journal of Psychopathology* **2023**, *29*, doi:10.36148/2284-0249-N284.

Navarro, E.; Goring, S.A.; Conway, A.R.A. The Relationship between Theory of Mind and Intelligence: A Formative g Approach. *J Intell* **2021**, *9*, doi:10.3390/jintelligence9010011.

Nguyen, T.N.; Gonzalez, C. Theory of Mind From Observation in Cognitive Models and Humans. *Top Cogn Sci* **2022**, *14*, doi:10.1111/tops.12553.

Patacchiola, M.; Cangelosi, A. A Developmental Cognitive Architecture for Trust and Theory of Mind in Humanoid Robots. *IEEE Trans Cybern* **2022**, *52*, doi:10.1109/TCYB.2020.3002892.

Patricio, M.L.M.; Jamshidnejad, A. Dynamic Mathematical Models of Theory of Mind for Socially Assistive Robots. *IEEE Access* **2023**, *11*, doi:10.1109/ACCESS.2023.3316603.

Premack, D.; Woodruff, G. Premack and Woodruff : Chimpanzee Theory of Mind. *Behavioral and Brain Sciences* **1978**, *4*.

Quesque, F.; Rossetti, Y. What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science* **2020**, *15*, doi:10.1177/1745691619896607.

Ross, S.; Pineau, J.; Chaib-Draa, B.; Kreitmann, P. Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. *Journal of Machine Learning Research* **2011**, *12*.

Ruiz-Serra, J.; Harré, M.S. Inverse Reinforcement Learning as the Algorithmic Basis for Theory of Mind: Current Methods and Open Problems. *Algorithms* 2023, *16*.

Schaafsma, S.M.; Pfaff, D.W.; Spunt, R.P.; Adolphs, R. Deconstructing and Reconstructing Theory of Mind. *Trends Cogn Sci* 2015, *19*.

Schurz, M.; Perner, J. An Evaluation of Neurocognitive Models of Theory of Mind. *Front Psychol* 2015, *6*.

Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; Tsvetkov, Y. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In Proceedings of the Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2023; Vol. 1.

Shain, C.; Paunov, A.; Chen, X.; Lipkin, B.; Fedorenko, E. No Evidence of Theory of Mind Reasoning in the Human Language Network. *Cerebral Cortex* **2023**, *33*, doi:10.1093/cercor/bhac505.

Stacy, S.; Gong, S.; Parab, A.; Zhao, M.; Jiang, K.; Gao, T. A Bayesian Theory of Mind Approach to Modeling Cooperation and Communication. *Wiley Interdiscip Rev Comput Stat* 2024, *16*.

Taber, K.S. Educational Psychology. In *Contemporary Trends and Issues in Science Education*; 2023; Vol. 56.

Vallacher, R.R.; Nowak, A.; Fennell, E. Mental Calibration: Fine Tuning the Dynamics of Mind and Action. In *Advances in Motivation Science*; 2023; Vol. 10.

van Duijn, M.; van Dijk, B.; Kouwenhoven, T.; de Valk, W.; Spruit, M.; van der Putten, P. Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Proceedings of the CoNLL 2023 - 27th Conference on Computational Natural Language Learning, Proceedings; 2023.

Vinitsky, E.; Lichtlé, N.; Yang, X.; Amos, B.; Foerster, J. Nocturne: A Scalable Driving Benchmark for Bringing Multi-Agent Learning One Step Closer to the Real World. In Proceedings of the Advances in Neural Information Processing Systems; 2022; Vol. 35.

Waade, P.T.; Enevoldsen, K.C.; Vermillet, A.Q.; Simonsen, A.; Fusaroli, R. Introducing Tomsup: Theory of Mind Simulations Using Python. *Behav Res Methods* **2023**, *55*, doi:10.3758/s13428-022-01827-2.

Wang, X.; Auyeung, B.; Pan, N.; Lin, L.Z.; Chen, Q.; Chen, J.J.; Liu, S.Y.; Dai, M.X.; Gong, J.H.; Li, X.H.; et al. Empathy, Theory of Mind, and Prosocial Behaviors in Autistic Children. *Front Psychiatry* **2022**, *13*, doi:10.3389/fpsyt.2022.844578.

Wäschle, M.; Thaler, F.; Berres, A.; Pölzlbauer, F.; Albers, A. A Review on AI Safety in Highly Automated Driving. *Front Artif Intell* 2022, *5*.

Wu, S.A.; Wang, R.E.; Evans, J.A.; Tenenbaum, J.B.; Parkes, D.C.; Kleiman-Weiner, M. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Top Cogn Sci* **2021**, *13*, doi:10.1111/tops.12525.

Zou, Z.; Mubin, O.; Alnajjar, F.; Ali, L. A Pilot Study of Measuring Emotional Response and Perception of LLM-Generated Questionnaire and Human-Generated Questionnaires. *Sci Rep* **2024**, *14*, doi:10.1038/s41598-024-53255-1.