Directions for computational theory of mind: Data, Metrics, Models and Mathematical Formalization

Prabhat Kumar, Erin Zaroukian, Douglas Summers-Stay, and Adrienne Raglin

US DEVCOM Army Research Laboratory, Adelphi, MD USA
prabhat.kumar.civ@army.mil

Abstract. This study expands on previous surveys of computational theory of mind (ToM) focusing on four key areas. Data: We attempt to characterize data needed for this research and propose creating procedurally generated, multi-modal synthetic data for training and testing ToM systems, addressing the lack of open-source data of agent behaviors in closed environments. Metrics: We explore ToM evaluation beyond the Sally-Anne Test, considering child development stages and natural language understanding as potential measures. Model: We investigate building on recent ToM models, exploring open-ended learning in reinforcement learning, and applying neuroscientific insights to model architecture. We also examine ToM applications in everyday technologies, leveraging state-of-the-art transformer technologies and multimodal datasets. Theoretical Formalization: We aim to bridge cognitive science and psychology concepts with mathematical approaches to facilitate algorithm development in ToM.

Keywords: theory of mind · game theory · multi-agent · machine-learning · artificial intelligence · intention · adversarial dynamics · computational · automation

1 Introduction

Computational theory of mind (ToM) is an area of artificial intelligence (AI) aiming to formalize and create algorithms for systems capable of inferring hidden internal states (intentions), and predict future behaviors and actions, of the agents it observes and interacts with. ToM has its foundations in cognitive science and psychology (see Wimmer and Perner [96], Premack and Woodruff [68]), but there are notable efforts to bring about its computational implementation, as we shall see. Korkmuz [43] mentions "ToM is a composite function, which involves memory, joint attention, complex perceptual recognition (such as face and gaze processing), language, executive functions (such as tracking of intentions and goals and moral reasoning), emotion processing-recognition, empathy, and imitation."

For many, such a capability is considered the holy grail of AI research having broad-reaching consequences in fields like social assistance (see Patricio [62],

Williams [95]), autonomous navigation Liu [51], video gaming in the creation of advanced characters to challenge players. Pijl [67] provides examples of behaviors requiring ToM which we present here along with some related references:

1. Intentionally communicating with others: Active communication to alter the listener's knowledge (see Baron-Cohen [7] within Corballis [13].)
2. Repairing failed communication: Recognizing action or dialogue may not make sense without context (see Bosco [9], Sidera [80]).
3. Teaching others: A teacher must recognize the understanding of their student to provide extra instruction as necessary (see Wellman [92], Knutsen [42]).
4. Intentionally persuading others: Altering another's beliefs.
5. Intentionally deceiving others: Specifically altering another's beliefs into a state of fallacy. (see Sarkadi [73], Alon [2]).
6. Building shared plans and goals: Understanding of another's perspective, knowledge and capabilities (see De Weerd [14]).
7. Intentionally sharing a focus or topic of attention: Understanding the perspective of another on a shared target (see Krych-Appelbaum [45], Buehler [11]).
8. Pretending: Not necessarily deception; in some cases all participants recognize the act (see Lillard [49]). Additionally, pretending requires a higher-order ToM to gauge another's beliefs about the pretender.

Contemporary computational ToM research has achieved notable results for single-agents in static environments (Rabinowitz [69], Raileanu [70], Nguyen [59]). Further the recent successes of transformer-based Large Language Models (LLMs) have prompted research into whether such models have various cognitive capabilities, including ToM. Aru [4] point out machines exploit "shortcuts", recognizing particular statistical features of the data (e.g. geometric arrangements) rather than inferring directly on an agent's "mindset." (This phenomena is seen earlier by Niven [60] in the context of natural language processing.) Kosinski [44] argues of the emergence of ToM in LLMs, and still others (Gandhi [22], Street [82], McDuff [55], Kennedy [39]) continue to argue and provide different perspective on the capabilities of these large generative models.

We address research in computational ToM by dividing the problem into four directions, as listed below:

- Data: Where we address the characteristics of data that have been used to date and how using multimodal data is essential for progress in the field.
- Metrics: As with measuring the visual reasoning capabilities of a computer-vision model or the "human-ness" of a natural language model, it is crucial to understand ToM usage to characterize the dimensions of inference a model can operate in.
- Models: Building on recent model studies by examining concepts in open-ended learning, neuroscience, and causal reasoning.
- Theory: Studies in cognitive development and psychology are obviously foundational here, but we hope to discuss a few mathematical ideas to facilitate algorithm development in the field.

As we consider these directions, we also consider viewing ToM from the lens of its conceptual and operational definitions which Baumeister [8] highlights: the conceptual definition includes the general abilities of inferring on hidden mental states like desires, goals or emotions, while the operational definition captures ToM "use" through characteristic performance on various tasks. Aru [4] points out that ToM cannot be wholistically captured through performance of individual tasks for it is the ability to perform AND adapt to a wide array of tasks and situations which enables the "true" use of ToM. For example, they advocate for open-ended learning (OEL) (see Hughes [33], Sigaud [81], DeepMind OEL Team [85]) to enable an agent to explore its environment and adapt to the various tasks and interactions it is presented with. This, of course, presents long-term consequences, in that given finite computational resources, developing and testing the full gamut of ToM characteristics and usage may not be viable, nor beneficial. Nonetheless, we aim to provide a foundation for a more dedicated and rigorous study of its application and usage.

2 Data

Widely available open-source data illustrating agent intention via behaviors in closed environments are in development. One source is Liu (2020) [51], where the goal was for an autonomous system to predict whether pedestrians were intending to cross. The study does not mention ToM explicitly, but their computer-vision-based model is an example for data fitting the operational definition of ToM. The dataset consists of "900 hours of driving scene videos of front, right, and left cameras, while the vehicle was driving in dense areas of five cities in the United States. The videos were annotated at 2fps with pedestrian bounding boxes and labels of crossing/not-crossing the street."

Gameplay datasets involving humans provide the most viable testing ground for ToM algorithms, as there is, at least assumed, intention ingrained in the play. The overall goal of any game is to win, which is broken down to the game objective: earn the most points, acquire the most territory, complete the most subtasks, etc. The game objectives inspire strategies; for example, focus on winning in tasks A,C,D, as B and E are difficult. Strategies are specialized into narrowed/directed intention; for example, distract opponent in a certain area on the board. Gameplay datasets involving artificial autonomous agents may be ingrained with "ToM-like structures" that went into its training. The following are examples for agent game play datasets and frameworks for generating game play, many of which follow from Tan [84]: Chess, Mitchell J [34]; Lichess Open Database [1]; Atari 2600, Kurin [47]; Super Mario Bros, Kauten [38]; Mincecraft, Guss [25]; StarCraft II, Vinyals [89].

We hypothesize that multiple modalities will be key in facilitating a model's understanding of the link between thought and physical action; linking what is said by, or described about, the agent, and what observable actions are performed. We use a simple thought-experiment; consider the following phrase: "I love it here!" One reader may not interpret the meaning behind this phrase

the same as another. Now suppose this was audibly stated by a human being. Audio presents pitch and tone data allowing us to infer who the speaker is. If it was from a child in a toy store, then one may infer genuine excitement, but if it were a physical laborer after they completed an arduous task, then their excitement could be questioned. The addition of another modality, audio, in form of vocal inflection, would aid in inferencing in this case. Further, the facts that the child is in a toy store or the laborer indeed completed an arduous task would not be apparent from either text or the audio modalities. Visual modality of each character in their respective environments, displaying body-language, further enhances inferencing. ToM traditionally deals with these three modalities (natural language, visual and audio); we have not come across any studies looking at other modalities (e.g. tactile) and even sub-modalities (e.g. infrared images, LiDAR point clouds) are limited. It is highly unlikely, if not impossible, for humans to perceive infrared without specific tools, so inferring on it makes no sense. What really happens is that information from these invisible regimes are transformed to be consumable by humans. For example, LiDAR data is processed until a map of the environment is constructed, or gravitational waves alter behavior of light which we can detect to process into signatures characterizing their origins. (A "competent" socially-intelligent agent would have the ability to recognize that their human partner would need these transformations to further infer on the traditionally non-interpretable information.) Di Vincenzo's dissertation [17] provides further insight into the multimodal nature of theory of mind, in particular as it relates to non-linguistic animals. Jin [36] presents one of the first multimodal ToM benchmarks, Multimodal Theory of Mind Question Answering (MMToM-QA) in the form of text descriptions along with series of images, as well as a novel architecture Bayesian Inverse Planning Accelerated by Language Models (BIP-ALM) to test this benchmark. Zhu [99] uses Simulation ToM to model beliefs during development of a common-ground between agents cooperating through multimodal interactions. Miniotaite [56] examines tabletop games Hanabi, Pandemic Hot Zone - Europe, Poker, and a custom game Peeker-Picker as opportunities for generating multimodal social data. Shi [77] expands on previous multimodal applications of ToM by incorporating multiple agents providing a pathway for systems tracking multiple individual behaviors, as well group dynamics.

In general, intention must be imbued within a dataset for a model to even consider it as a subject of inference. In addition to curating human gameplay, we advocate for generating one's own agent behavior datasets through resources like Farama Foundation [12], [86], NetLogo [94], or any game engine allowing for reinforcement learning (RL) plugins. The benefits of using agents trained via RL are that the algorithm's parameters and learned policies are quantified and available as ground-truth for comparisons with a ToM model's inferences, which will further allow for creation of elusive metrics.

3 Metrics

How do we "measure" ToM? A classical evaluation of ToM capabilities is the Sally-Anne Test, which tests for explicit knowledge of a false belief. Some argue, however, that a more implicit form of ToM may be present in humans, other animals, or perhaps even computational models that lack the language, executive function, and neural development to succeed at an explicit Sally-Anne Test; see Rakoczy [71]. We hypothesize as Aru [4] does, that ToM is a process not tied to performing any one task is particular; it also requires adaptability in learning and understanding.

Attempting the ascertain the ToM capabilities of LLMs is quite a popular subject. Summers-Stay [83] implemented tests from Kaland [37] on GPT-3 and found GPT-3 was able to pass all of these tests, but it was very inconsistent in its abilities to answer questions generally. Current top-end LLMs have no problem with these kinds of tasks if prompted appropriately. Xu [97] address the lack of personality traits, preferences and motivations in human ToM tests used for LLMs by introducing OpenToM, a human-in-the-loop generated benchmark providing for these shortcomings while also assessing the model's capacity for understanding both physical and psychological worlds. They further attempt to mitigate spurious correlations, a challenge raised in Aru [4], by manually revising narratives with "substantial lexical overlap with questions or those that provide shortcuts for answering them." Further they employ concepts of causal reasoning based on Judea Pearl's works (see [64], [65], [66]) to highlight spurious cues. Other recent examples of LLM ToM benchmarks include FANToM from Kim [41], HI-TOM from He [27], ToMATO from Shinoda [78], EgoSocialArena from Hou [31] and Le [48].

Another recent attempt at measuring ToM in computational systems is the AGENT benchmark from Shu [79] which creates a dataset consisting of videos of agents performing a series of four core psychological reasoning tasks which they use to compare the performances of two ToM models, one based on Bayesian Inferencing Planning (BIPACK) and another based on a neural network architecture, ToMnet-G. They go on to discuss, "In addition to this minimal set of concepts, a model may also need to understand other concepts to pass a full battery of core intuitive psychology, including perceptual access and intuitive physics. Although this minimal set does not include other concepts of intuitive psychology such as false belief, it is considered part of 'core psychology' in young children who cannot yet pass false belief tasks, and forms the building blocks for later concepts like false belief."

Hagendorff [26] offers prescriptive insight introducing methods for testing and interpreting behaviors of LLMs. Sclar (2024) [76] "introduce ExploreToM, an A*-powered algorithm [leveraging LLMs like Llama-3.1-70B, GPT-4o and Mixtral-8x7B-Instruct] for generating reliable, diverse, and challenging theory of mind data that can be effectively employed for testing or fine-tuning LLMs." Sclar (2022) [75] provides a situated, multi-agent environment, SymmToM, incorporating ideas from reinforcement learning to test their capabilities.

We propose considering research in child development as a framework for developing a "battery" of tests for measuring ToM capabilities. Specifically, the stages, as given in Wellman and Liu [93] and summarized in Baumeister [8]:

1. Diverse Desires: Recognizing two agents have different DESIRES about the same object. (e.g. Abby wants the coffee, but Mike does not.)
2. Diverse Beliefs: Recognizing two agents having different BELIEFS about the same object. (e.g. Abby thinks the coffee is bitter, Mike thinks its sweet.)
3. Knowledge Access: Ability to judge knowledge of another agent not sharing the participants knowledge (e.g. Mike realizes Abby knows how to play a certain video game that he does not.)
4. Contents False Belief: Recognizing an individual's false beliefs about a container's contents. (e.g. Abby told Mike she thinks there is a pizza in his lunch bag, when he actually packed a sandwich.)
5. Explicit False Belief: Predicting subsequent behavior of another individual with a false-belief. (e.g. Abby asks Mike to trade his "pizza" for her sandwich.)
6. Belief Emotion: Judging how another individual feels based on a false belief. (e.g. Abby is excited about the "pizza" she's about to get.)
7. Real-Apparent Emotion: Recognizing an individual may feel a certain way, but display different emotions. (e.g. Abby looked sad when she actually traded a sandwich for another sandwich, instead of a pizza, but she was actually happy because her second sandwich was not dry like the first.)

Baumeister [8] goes on to comment about "higher-order reasoning", as well. That is, understanding an agent may have a false-belief about another agent's beliefs. Recursive thinking (Raileanu [70], Devaine [16]) is an example of this.

While all of these can be tested through natural language modalities, in connection to our goal of promoting multimodal explorations of ToM, we propose generalizing these to, say, complimentary visual modalities, as well. An example for the Diverse Desires (and possibly even Diverse Beliefs) could be illustrated using a (series of) videos depicting two agents and an object, where one agent proceeds towards the object and another retreats from it. Knowledge-access, as another example, can be illustrated through a video depicting an agent methodically completing a task unknown to the model. Strict audio modality applications of these tasks can be accomplished, for example, through verbal story-telling or engagement in dialogue. Again, we emphasize the use of multiple modalities as a mechanism for uncovering intention. An agent seemingly behaving randomly to the naked eye, using RGB data, may actually act based on the interpretation of say LiDAR point-cloud data illustrating the presence of objects of interest in the environment.

One important issue is that most tests of theory of mind were designed for humans. If a human can pass such a test, we can conclude that they have other theory-of-mind abilities as well. However, it is less clear what the ability to pass such a test implies for an LLM. Even if it can predict what someone else will think when taking such a test, will it use this ability when, for example,

teaching a new concept? Does the LLM's attention (or probability weighting; Kosinski [44]) match the patterns of attention/eye gaze expected from a human in an implicit test of ToM, such as an anticipatory looking test of a violation of expectation test?

4 Models

How do we build on recent studies developing ToM models? Can neuroscientific research point us towards properly designing model architectures, as in the case of the development of convolutional neural networks? How do we characterize problems from the perspective of fundamental machine learning? How does work in causal reasoning relate to ToM?

Nebreda [57] categorizes ToM models into three particular types: Cognitive, black-box and bio-inspired. Bio-inspired models constitute those based on neuroscientific study; Nebreda [57] references Ask [5] among others which discuss the difficulties of computationally modeling biology and advocates for multi-level modeling as an approach as opposed to a single model capturing all neurobiological phenomena. In terms of pure biology and neuroscience, Saxe (2006) [74] compiles neuroscientific ToM research through 2006 which mentions the recruitment of the right temporo-parietal junction (RTPJ) in reasoning about others mental states, in particular, "the RTPJ does appear to reflect the functioning of a specialized, domain-specific mechanism for reasoning about beliefs." Wade [90] tests hypotheses of the interplay between ToM and executive function (EF) from the perspective of neurological development, cites "the importance of the superior temporal regions" based on Apperly [3].

Gallese [21] discusses a class of neurons, mirror neurons, discovered at the time, and posits their use in the "action-execution/observation matching system" used for "mind-reading." Keysers [40] proposes Hebbian learning to explain the existence of mirror neurons citing Hebb [28], "'When an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place [...] such that A's efficiency, as one of the cells firing B, is increased'. Put in simpler words: 'neurons that fire together wire together'." More recently Mohammadi [24] assembles these ideas into a machine learning model for mirror neurons and implements a ToM experiment using the River Raid Atari game environment by OpenAI [10].

Computational ToM problems can be posed from the lens of (un-/semi-)supervised and reinforcement learning. With supervised learning, the goal of a ToM model is to characterize, and predict the behaviors of, agents based on ground-truth; such an approach is implemented in Rabinowitz [69], for example. Supervised learning in this manner can limit a model's generalizability as it becomes a task of exhaustively expressing various behaviors and actions. Hewson [29] uses a self-supervised approach with ToM concepts tying "extrinsic motivations, such as [reinforcement learning from human feedback]" with "intrinsic motivations" that achieve its own goals, in order to facilitate model understanding of human desires.

Unsupervised learning for ToM systems could provide insight into cognitive behaviors/structures not characterized before. A simple example consists of several agents with uncharacterized traits; it would be up to the model to group their behaviors accordingly and up to the researcher to define these grouping. We discussed RL before from the perspective of generative agent behavior data; in terms of a ToM model it allows for illustration of the operational definition of ToM, but defining proper reward functions encouraging the model to reason on another agent's hidden states is difficult (Aru [4]), but has been attempted (Oguntola [61]).

Earlier, we noted ideas of causal reasoning being used to infer on spurious cues in the works of Xu [97]; see works of Fears [20], Rawal [72] for other examples in the use of causality. Causal models form another approach to ToM representation. Ho [30] describes ToM as a causal model, especially when viewed from the perspective of planning. Lombard [52] argues ToM also aims to understand that, "actions based on such understanding [of emotions, attention, desires, beliefs] have causes and effects" and goes on to analyze ToM by order as described in Dennett [15]:

1. Zero-order ToM ascribes no mentality to an individual, but assumes that behavior of the individual is governed by instincts, reflexes, or conditioning.
2. First-order ToM attributes emotions, attention, desires, intentions, or beliefs to the individual and that some forms of behaviors are governed by these entities. This level, however, presumes no understanding of the minds of other individuals.
3. Second-order ToM requires an individual to attribute a ToM to other individuals and to use this in their understanding of the behavior of others.
4. Third-order ToM requires an individual A to attribute to a second individual B an understanding of the ToM of A.
5. Higher orders of ToM require an individual to represent at least two mental states, their own and that of someone else.

Delineating and characterizing model order provides insight into its capabilities. In our earlier survey (Kumar [46]) we provide for another delineation based on model perspective: third-person versus first-person.

As mentioned previously, LLMs serve as viable experimental subjects in themselves for testing ToM abilities due to the strong link between language and ToM (see) even if the debates as to their actual capabilities have not been settled (see Kosinski [44], Ullman [87], Zhou [98], Hou [32]) The rapid growth of these technologies provides hope in the development of ToM faculties that incorporate the above ideas, and each new generation lends itself to more rigorous scrutiny.

5 Theory

While cognitive science and psychology serve as the theoretical foundation for ToM research, we examine mathematical perspectives to facilitate algorithm development.

Baker [6] describes humans' understanding of the internal states of others based on observable actions using the framework of Bayesian Inverse Planning (BIP), which serves as a foundation for a number of contemporary studies in computational ToM. The basis of the framework lies within modeling agent behaviors and the associated uncertainties within closed environments using Markov Decision Processes (MDPs).

MDPs are defined (see Uther [88]) as a tuple $\{S, A, p, r\}$, where

- S is the state space: Space of possible configurations of the environments containing the agent;
- A is the action space: Space of possible actions of the agent within the environment;
- p is the transition function: Function representing the probability of a $s' \in S$ given another state $s \in S$ and action $a \in A$;
- r is the reward function: Function representing the reward (punishment) the agent received for taking an action a ($\in A$) in state s ($\in S$).

MDPs can be generalized to Partially Observable MDPs, POMDPs, (Uther [88]) if we assume the agent cannot have complete knowledge of the environment, which is consistent with reality; such a model has been used in, for example, Rabinowitz [69] to model agent behavior in a closed environment which is used to train a neural network to characterize and predict future behaviors.) Baker [6] discusses the BIP model in terms of Environment (Env), encoded as S in an MDP, Action, encoded as A, and Goal (encoded in r). Further, they formalize probabilistic planning, then, as P(Action|Goal,Env) where P denotes a probability, from which BIP is given by Bayes' Rule:

$$P(\text{Goal}|\text{Action,Env}) \propto P(\text{Action}|\text{Goal,Env})P(\text{Goal}|\text{Env})$$

Jara-Ettinger [35] describes ToM in terms of inverse reinforcement learning (IRL), which from Ng [58] is formalized as determining a reward function based on observed behaviors, sensory and environmental inputs; Ng also develops the IRL problem in terms of a MDPs.

(PO)MDPs serve as a veritable experimental ground for testing these frameworks. An interesting extension would be including a notion of "indirect information"; that is, information provided to the agent that alters their behavior but is not perceivable through direct observation. For example, observing an agent change course not because of any obstacle on their path or the sudden appearance of a new goal item within the environment, but due to information they may have received externally. The cause of the change could be "invisible" based on the observer's perceptive capabilities (e.g. an agent acting on LiDAR data while their observer only has access to RGB). The situation where an observer cannot access secret communications between, say, a subject agent and a third-party is subsumed into that where the observer does not have the capability for such access. The goals of this problem then become (1) recognizing "indirect information" as a, now, measurable cause and (2) identifying the actual source of the "indirect information". In other words, ascertaining an explanation provided the

given observations; this is the basis for abductive reasoning (see Douven [19]). We can think of ToM as a special application of abductive reasoning; using what we can perceive about an agent, how do we explain their behavior? Gordon [23] provides a computational approach to abductive reasoning which uses a knowledge-base of pre-determined (joint/conditional) probabilities of various observations provided certain hypotheses. Using this knowledge, their Etcetera-Abduction system performs a combinatorial search of potential explanations for a given set of observations. That is, it essentially solves

$$\text{argmax}_{H\in\mathcal{H}}\text{eval}(H) = P(H|O) = \frac{P(O|H)P(H)}{P(O)},$$

where $\mathcal{H}$ gives the space of potential hypotheses, eval is a function used to evaluate candidate hypotheses for minimizing their cost as related to explaining observation, O. The trickiest task in using this model is formalizing an extensive knowledge-base and quantifying the associated event probabilities. The problem becomes more intractable as we consider deeper causal chains possible for the agent and longer observer context windows (i.e. how far back in its memories and experiences does it have go to interpret a set of events?) However, such approach allows for explanations involving potential unknown actors and causes, mostly as lower probability explanations, for given situations.

Another approach for computational ToM individualizes the model to the type of observer. Patricio [62] uses the idea of fuzzy cognitive maps (FCMs). To summarize the mathematical framework as the authors present, the evolution of a system is provided by variables called concepts; C_i being the i-th concept (e.g. emotion). A representation A_i of the i-th concept provides a possible instantiation (e.g. happiness). $\mathbb{C}$ is the set of all concepts, $\mathbb{A}$ is the set of all instantiations. Concepts can be linked to other concepts and the weight of the links determines their influence; such weights can vary as functions of concepts and their representations. The set of all simple links is $\mathbb{L}$; those connecting two concepts not connected to any third. $\bar{\mathbb{L}}$ is the set of all complex links: simple links and that of a third concept affecting it. Patricio delineates the dynamic equation for updating concepts over time:

$$C_j(k+1) = h\Big(\sum_{\forall i|(i,j)\in\mathbb{L}} f(C_i(k),C_j(k))C_i(k) + \sum_{\forall i;\exists l|(i,j,l)\in\bar{\mathbb{L}}} g(C_l(k),C_i(k),C_j(k))C_i(k)+\alpha_j C_j(k)\Big),$$

where h is a threshold function constraining C_j, $f:\mathbb{A}^2\longrightarrow[-1,1]$, $g:\mathbb{A}^3\longrightarrow[-1,1]$, α_j correspond to the influence C_j, realized at the current timestep, has on the same concept during the next step. Further, the authors personalized the weights of the links between concepts using a loss optimization strategy involving quantifications of the individual's linguistic responses to a survey about their "preferences, rationally perceived knowledge, and general world knowledge."

One more example of individualization, albeit in a more group-like manner, is Diaconescu [18] applying a Hierarchical Gaussian Filter (HGF) (see Mathys (2011) [53] and Mathys (2014) [54]) for tracking shifting intentions and the associated volatilities. They describe that "an agent uses a sequence of sensory inputs

to make inferences on a hierarchy of hidden states, $x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}$ (where k is the trial index and n is the number of levels in the hierarchy)." In their framework, x_1 represents a binary variable representing belief about the accuracy [0 or 1] of advice provided by another actor. This variable depends on x_2, representing "the belief about the adviser's tendency to deliver accurate advice", which in turn depends on x_3, the "volatility of the adviser's intentions"; the latter two evolve as Gaussian random walks, and represent beliefs about advice accuracy. They develop the following generative model for their HGF implementation:

$$p\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_2^{(k-1)}, x_3^{(k-1)} | \kappa, \omega, \varphi\right) =$$
$$p\left(x_1^{(k)} | x_2^{(k)}\right) p\left(x_2^{(k)} | x_2^{(k-1)}, x_3^{(k)}, \kappa, \omega\right) p\left(x_3^{(k)} | x_3^{(k-1)}, \varphi\right) p\left(x_2^{(k-1)}, x_3^{(k-1)}\right)$$

where κ represents the coupling between x_2 and x_3, ω is the "tonic component of the log-volatility at the second level", and φ denotes the evolutionary rate of x_3. This HGF model was used to describe participants' learning of their corresponding actor's intentions and the parameters used be associated with different behaviors, employed strategies and tendencies.

While we can think of ToM as an ability and suggest individual models for it, in hopes of raising discussion to bring generalization to the concept, we introduce a different perspective treating ToM as a map. Consider an observer O and a subject, S. Specifically, O has a function,

$$T_{O,S} : C \longrightarrow F,$$

mapping from a current state, C, of S to its future state, F. Thinking of ToM fundamentally, the past experiences of O influence its characterizations on the subject; see Rabinowitz [69] for example. So we can refine the map:

$$T_{O,S} : C \times E \longrightarrow F,$$

where E represents the past experiences of O.

We could break down $T_{O,S}$: One could argue that past experiences, or at least the way they are perceived, are shaped by various characteristics of the individual observer. In this manner, we can think of E as an output to another map,

$$P_\theta : I \longrightarrow E,$$

where I represents physical inputs; P is initially parametrized by a quantification, θ, of the observer's tendencies, temperament, biases, etc. For AI models, these concepts would most certainly depend on their training. Training sets imbued with bias or specialized on certain data/tasks will result in differing θ values. This quantification can be, and is, the basis of research in the cognitive sciences and psychology, (see Diaconescu [18], Patricio [63] and [62]). This seems to lead us in a circle: Doesn't this mean we need a characterization of the observer's ToM before we can use them to model others? Perhaps a model of this nature may lead to development of a series of models each with differing perspectives that could act in a collaborative way. Intuitively, for any two human

observers, one does not have the same ToM as the other about a subject, which we can formalize:

$$\forall i \text{ consider } O_i \text{ with ToM function } T_{O_i}, \text{ we have } T_{O_n} \neq T_{O_m} \text{ for } n \neq m.$$

These differences in T_{O_n} and T_{O_m} can be defined in P_{θ_n} and P_{θ_m}, respectively. This particular uniqueness endorses collaboration in humans ("two heads are better than one") and immediately gives rise to a concept for adversarial interaction. There is, of course, a concept for neutral (neither cooperative nor adversarial) engagements, as well.

Another angle considers a ToM function as a composition. One example:

$$T_1 : C \times E \longrightarrow R \tag{1}$$

$$T_2 : R \longrightarrow F, \tag{2}$$

where R is an (intermediate) characterization of a subject based on its past and current behaviors.

There are likely several decompositions of T, but this perspective allows us to consider T as a collection of functions each responsible for various ToM tasks; allowing us the flexibility of refining several sub-models that work in conjunction with one another. ToM using solely visual input, intuitively, uses different faculties than that which uses solely audible inputs. Recognition that some visual and audio inputs may be linked provides synergistic inferencing capabilities, which we alluded to above with using "series" of ToM models. The difference is using several ToM models with differing θ (e.g. two people reasoning on the same social phenomena) versus a one model capturing multiple ToM sub-abilities (e.g. one person reasoning on two different social phenomena); of course, there's nothing disallowing mingling of these two concepts.

Viewing computational ToM through this generalized perspective allows for further concepts, like time evolution. We slowly push towards a computational concept for ToM that accounts for multiple modalities, multiple agents and shifting environments, but we must also consider how the dynamics of these elements shape ToM reasoning; after all, ToM considers past experiences, so how do current experiences transition to ingrained knowledge of a model; that is, how do we ensure models are continuously learning and evolving? (See the following for research in AI continual learning: Wang [91], Liu (2017) [50].) The introduction of a time parameter for T can help conceptualize, but specific implementation needs careful treatment. As a model continuously learns we can further hold that, similar to two agents espousing different ToM models, that for any one agent, a mental model at one timestep may not necessarily be identical to that of another timestep; that is, given a particular observer O with ToM function $T_{O,t}$ at a particular time t, we hold that

$$\forall \text{ timesteps } t_i, \exists \text{ a timestep } t_j \ (i < j) \text{ such that } T_{O,t_i} \neq T_{O,t_j}.$$

That is, we hold that a model must evolve after a certain point. We leave open discussions on whether it makes sense for a model to be held constant in certain circumstances, and, broadly, how to continue developing these mathematical ideas.

6 Discussion & Conclusion

In this paper, we attempt to provide additional perspective for computational ToM by exploring research through four directions: (1) data, (2) metrics, (3) models, and (4) mathematical formalizations. The ideas in this paper most certainly lend themselves to further exposition and rigor and we welcome such discussions. Finding data to reliably train models remains a generic research challenge; we can mitigate these challenges through research into generative technologies. Measuring ToM usage is reduced to measuring performance of computational systems on a selection of concrete tasks associated with ToM abilities. Adaptability to various tasks is key, be it through enhancements in transfer learning, applications of meta-learning, etc. One discussion we hope to address later is research comparing ToM acquisition and usage from pre-/non-verbal humans to those with verbal capabilities and how it allows discussion into ToM capabilities of multimodal models. Just as biology inspired research and design of convolutional neural networks in computer vision, and other artificial capabilities, we discussed similar biologically-inspired pathways for developing ToM faculties. We mentioned the possibility of a multimodal"aware" model when providing an example for explaining behaviors of an agent acting within a regime not accessible to the observer. One question we hope to discuss further is would the observer's realization of extraneous regimes fall into ToM phenomena or is it governed by another? We aimed to open up discussion about the similarities and differences between mathematical models of ToM, as well as provide the initial seeds to generalize some concepts to provide further perspective on tackling research in this field.

References

1. Lichess.org open database, https://database.lichess.org/
2. Alon, N., Schulz, L., Rosenschein, J.S., Dayan, P.: A (dis-) information theory of revealed and unrevealed preferences: emerging deception and skepticism via theory of mind. Open Mind **7**, 608–624 (2023)
3. Apperly, I.A., Samson, D., Chiavarino, C., Humphreys, G.W.: Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. Journal of cognitive neuroscience **16**(10), 1773–1784 (2004)
4. Aru, J., Labash, A., Corcoll, O., Vicente, R.: Mind the gap: challenges of deep learning approaches to theory of mind. Artificial Intelligence Review pp. 1–16 (2023)
5. Ask, M., Reza, M.: Computational models in neuroscience: how real are they? a critical review of status and suggestions. Austin Neurology & Neurosciences **1**(2), 1008 (2016)
6. Baker, C.L., Saxe, R., Tenenbaum, J.B.: Action understanding as inverse planning. Cognition **113**(3), 329–349 (2009)
7. Baron-Cohen, S.: The evolution of a theory of mind. Oxford University Press (1999)
8. Baumeister, F., Wolfer, P., Sahbaz, S., Rudelli, N., Capallera, M., Daum, M.M., Samson, A.C., Corrigan, G., Naigles, L., Durrleman, S.: Measuring theory of mind: a preliminary analysis of a novel linguistically

simple and tablet-based measure for children. Frontiers in Developmental Psychology **2** (2024). https://doi.org/10.3389/fdpys.2024.1445406, https://www.frontiersin.org/journals/developmental-psychology/articles/10.3389/fdpys.2024.1445406
9. Bosco, F.M., Gabbatore, I.: Theory of mind in recognizing and recovering communicative failures. Applied Psycholinguistics **38**(1), 57–88 (2017)
10. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym (2016), https://arxiv.org/abs/1606.01540
11. Buehler, M.C., Weisswange, T.H.: Theory of mind based communication for human agent cooperation. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS). pp. 1–6 (2020). https://doi.org/10.1109/ICHMS49158.2020.9209472
12. Chevalier-Boisvert, M., Dai, B., Towers, M., de Lazcano, R., Willems, L., Lahlou, S., Pal, S., Castro, P.S., Terry, J.: Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. CoRR **abs/2306.13831** (2023)
13. Corballis, M.C., Lea, S.E.: The descent of mind: Psychological perspectives on hominid evolution. Oxford University Press (1999)
14. De Weerd, H., Verbrugge, R., Verheij, B.: Higher-order theory of mind in the tacit communication game. Biologically Inspired Cognitive Architectures **11**, 10–21 (2015)
15. Dennett, D.C.: The intentional stance. MIT press (1989)
16. Devaine, M., Hollard, G., Daunizeau, J.: Theory of mind: Did evolution fool us? PLOS ONE **9**(2), 1–12 (02 2014). https://doi.org/10.1371/journal.pone.0087619, https://doi.org/10.1371/journal.pone.0087619
17. Di Vincenzo, L., et al.: Theory of mind in non-linguistic animals: a multimodal approach (2024)
18. Diaconescu, A.O., Mathys, C., Weber, L.A.E., Daunizeau, J., Kasper, L., Lomakina, E.I., Fehr, E., Stephan, K.E.: Inferring on the intentions of others by hierarchical bayesian learning. PLOS Computational Biology **10**(9), 1–19 (09 2014). https://doi.org/10.1371/journal.pcbi.1003810, https://doi.org/10.1371/journal.pcbi.1003810
19. Douven, I.: Abduction. https://plato.stanford.edu/archives/sum2021/entries/abduction/ (2021)
20. Fears, A., Raglin, A., Basak, A.: Causal intervention and semantic knowledge for object relationships. In: 2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI). pp. 17–22. IEEE (2024)
21. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends in cognitive sciences **2**(12), 493–501 (1998)
22. Gandhi, K., Fränken, J.P., Gerstenberg, T., Goodman, N.D.: Understanding social reasoning in language models with language models (2023), https://arxiv.org/abs/2306.15448
23. Gordon, A.S., Feng, A.: Searching for the most probable combination of class labels using etcetera abduction. In: 2023 57th Annual Conference on Information Sciences and Systems (CISS). pp. 1–6. IEEE (2023)
24. Gorgan Mohammadi, A., Ganjtabesh, M.: On computational models of theory of mind and the imitative reinforcement learning in spiking neural networks. Scientific Reports **14**(1), 1945 (2024)
25. Guss, W.H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M., Salakhutdinov, R.: Minerl: A large-scale dataset of minecraft demonstrations (2019), https://arxiv.org/abs/1907.13440

26. Hagendorff, T.: Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988 **1** (2023)
27. He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., Deng, N.: Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models (2023), https://arxiv.org/abs/2310.16755
28. Hebb, D.O.: The organization of behavior: A neuropsychological theory. Psychology press (2005)
29. Hewson, J.T.S.: Combining theory of mind and kindness for self-supervised human-ai alignment (2024), https://arxiv.org/abs/2411.04127
30. Ho, M.K., Saxe, R., Cushman, F.: Planning with theory of mind. Trends in Cognitive Sciences **26**(11), 959–971 (2022)
31. Hou, G., Zhang, W., Shen, Y., Tan, Z., Shen, S., Lu, W.: Egosocialarena: Benchmarking the social intelligence of large language models from a first-person perspective (2025), https://arxiv.org/abs/2410.06195
32. Hou, G., Zhang, W., Shen, Y., Wu, L., Lu, W.: Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. arXiv preprint arXiv:2407.01455 (2024)
33. Hughes, E., Dennis, M., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., Rocktaschel, T.: Open-endedness is essential for artificial superhuman intelligence (2024), https://arxiv.org/abs/2406.04268
34. J, M.:
35. Jara-Ettinger, J.: Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences **29**, 105–110 (2019). https://doi.org/https://doi.org/10.1016/j.cobeha.2019.04.010, https://www.sciencedirect.com/science/article/pii/S2352154618302055, artificial Intelligence
36. Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J.B., Shu, T.: Mmtom-qa: Multimodal theory of mind question answering (2024)
37. Kaland, N., Møller-Nielsen, A., Smith, L., Mortensen, E.L., Callesen, K., Gottlieb, D.: The strange stories test: A replication study of children and adolescents with asperger syndrome. European child & adolescent psychiatry **14**, 73–82 (2005)
38. Kauten, C.: Super Mario Bros for OpenAI Gym. GitHub (2018), https://github.com/Kautenja/gym-super-mario-bros
39. Kennedy, S.M., Nowak, R.D.: Cognitive flexibility of large language models. In: ICML 2024 Workshop on LLMs and Cognition (2024)
40. Keysers, C., Perrett, D.I.: Demystifying social cognition: a hebbian perspective. Trends in cognitive sciences **8**(11), 501–507 (2004)
41. Kim, H., Sclar, M., Zhou, X., Bras, R.L., Kim, G., Choi, Y., Sap, M.: Fantom: A benchmark for stress-testing machine theory of mind in interactions (2023), https://arxiv.org/abs/2310.15421
42. Knutsen, J., Frye, D., Sobel, D.M.: Theory of learning, theory of teaching, and theory of mind. O. Saracho & Spodek (Eds.), Contemporary perspectives on early childhood education pp. 269–290 (2014)
43. Korkmaz, B.: Theory of mind and neurodevelopmental disorders of childhood. Pediatric research **69**(8), 101–108 (2011)
44. Kosinski, M.: Theory of mind may have spotaneously emerged in large language models (11 2023). https://doi.org/https://doi.org/10.48550/arXiv.2302.02083

45. Krych-Appelbaum, M., Law, J.B., Jones, D., Barnacz, A., Johnson, A., Keenan, J.P.: "i think i know what you mean": The role of theory of mind in collaborative communication. Interaction Studies **8**(2), 267–280 (2007)
46. Kumar, P., Raglin, A., Richardson, J.: Surveying computational theory of mind and a potential multi-agent approach. In: International Conference on Human-Computer Interaction. pp. 376–390. Springer (2024)
47. Kurin, V., Nowozin, S., Hofmann, K., Beyer, L., Leibe, B.: The atari grand challenge dataset. arXiv preprint arXiv:1705.10998 (2017)
48. Le, M., Boureau, Y.L., Nickel, M.: Revisiting the evaluation of theory of mind through question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5872–5877 (2019)
49. Lillard, A.S.: Pretend play skills and the child's theory of mind. Child development **64**(2), 348–371 (1993)
50. Liu, B.: Lifelong machine learning: a paradigm for continuous learning. Frontiers of Computer Science **11**(3), 359–361 (2017)
51. Liu, B., Adeli, E., Cao, Z., Lee, K.H., Shenoi, A., Gaidon, A., Niebles, J.C.: Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robotics and Automation Letters **5**(2), 3485–3492 (2020)
52. Lombard, M., Gärdenfors, P.: Causal cognition and theory of mind in evolutionary cognitive archaeology. Biological Theory **18**(4), 234–252 (2023)
53. Mathys, C., Daunizeau, J., Friston, K.J., Stephan, K.E.: A bayesian foundation for individual learning under uncertainty. Frontiers in human neuroscience **5**, 39 (2011)
54. Mathys, C.D., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., Stephan, K.E.: Uncertainty in perception and the hierarchical gaussian filter. Frontiers in human neuroscience **8**, 825 (2014)
55. McDuff, D., Munday, D., Liu, X., Galatzer-Levy, I.: Cognitive assessment of language models. In: ICML 2024 Workshop on LLMs and Cognition (2024)
56. Miniotaite, J., Pereira, A.: Tabletop games as multimodal datasets for social ai
57. Nebreda, A., Shpakivska-Bilan, D., Camara, C., Susi, G.: The social machine: artificial intelligence (ai) approaches to theory of mind. In: The theory of mind under scrutiny: psychopathology, neuroscience, philosophy of mind and artificial intelligence. pp. 681–722. Springer (2024)
58. Ng, A.Y., Russell, S.: Algorithms for inverse reinforcement learning. In: ICML. vol. 1, p. 2 (2000)
59. Nguyen, T.N., Gonzalez, C.: Theory of mind from observation in cognitive models and humans. Topics in Cognitive Science **14**(4), 665–686 (2022). https://doi.org/https://doi.org/10.1111/tops.12553, https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12553
60. Niven, T., Kao, H.Y.: Probing neural network comprehension of natural language arguments (2019), https://arxiv.org/abs/1907.07355
61. Oguntola, I., Campbell, J., Stepputtis, S., Sycara, K.: Theory of mind as intrinsic motivation for multi-agent reinforcement learning. arXiv preprint arXiv:2307.01158 (2023)
62. Patrício, M., Jamshidnejad, A.: Mathematical models of theory of mind (09 2022). https://doi.org/10.48550/arXiv.2209.14450
63. Patrício, M.L.M., Jamshidnejad, A.: Dynamic mathematical models of theory of mind for socially assistive robots. IEEE Access **11**, 103956–103975 (2023). https://doi.org/10.1109/ACCESS.2023.3316603

64. Pearl, J.: Causal inference in statistics: An overview (2009)
65. Pearl, J.: Causality. Cambridge university press (2009)
66. Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Communications of the ACM **62**(3), 54–60 (2019)
67. Pijl, L.: Modelling the evolution of theory of mind. Ph.D. thesis, Faculty of Science and Engineering (2011)
68. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978). https://doi.org/10.1017/S0140525X00076512
69. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.A., Botvinick, M.: Machine theory of mind. In: International conference on machine learning. pp. 4218–4227. PMLR (2018)
70. Raileanu, R., Denton, E., Szlam, A., Fergus, R.: Modeling others using oneself in multi-agent reinforcement learning. In: Krause, A., Dy, J. (eds.) 35th International Conference on Machine Learning, ICML 2018. pp. 6779–6788 (2018)
71. Rakoczy, H.: Foundations of theory of mind and its development in early childhood. Nature Reviews Psychology **1**(4), 223–235 (2022)
72. Rawal, A., Raglin, A., Rawat, D.B., Sadler, B.M., McCoy, J.: Causality for trustworthy artificial intelligence: Status, challenges and perspectives. ACM Computing Surveys (2024)
73. Sarkadi, S., Panisson, A., Bordini, R., McBurney, P., Parsons, S., Chapman, M.: Modelling deception using theory of mind in multi-agent systems. AI COMMUNICATIONS **32**(4), 287–302 (2019). https://doi.org/10.3233/AIC-190615
74. Saxe, R., Baron-Cohen, S.: Editorial: The neuroscience of theory of mind. Social Neuroscience **1**(3-4), 1–9 (2006). https://doi.org/10.1080/17470910601117463, https://doi.org/10.1080/17470910601117463, pMID: 18633771
75. Sclar, M., Neubig, G., Bisk, Y.: Symmetric machine theory of mind. In: International Conference on Machine Learning. pp. 19450–19466. PMLR (2022)
76. Sclar, M., Yu, J., Fazel-Zarandi, M., Tsvetkov, Y., Bisk, Y., Choi, Y., Celikyilmaz, A.: Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. arXiv preprint arXiv:2412.12175 (2024)
77. Shi, H., Ye, S., Fang, X., Jin, C., Isik, L., Kuo, Y.L., Shu, T.: Muma-tom: Multimodal multi-agent theory of mind (2025), https://arxiv.org/abs/2408.12574
78. Shinoda, K., Hojo, N., Nishida, K., Mizuno, S., Suzuki, K., Masumura, R., Sugiyama, H., Saito, K.: Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind (2025), https://arxiv.org/abs/2501.08838
79. Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J., Ullman, T.: Agent: A benchmark for core psychological reasoning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 9614–9625. PMLR (18–24 Jul 2021)
80. Sidera, F., Perpiñà, G., Serrano, J., Rostan, C.: Why is theory of mind important for referential communication? Current Psychology **37**, 82–97 (2018)
81. Sigaud, O., Baldassarre, G., Colas, C., Doncieux, S., Duro, R., Perrin-Gilbert, N., Santucci, V.G.: A definition of open-ended learning problems for goal-conditioned agents (2023)
82. Street, W., Siy, J.O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Dunbar, R.I., et al.: Llms achieve adult human performance on higher-order theory of mind tasks. arXiv preprint arXiv:2405.18870 (2024)

83. Summers-Stay, D., Bonial, C., Voss, C.: What can a generative language model answer about a passage? In: Proceedings of the 3rd Workshop on Machine Reading for Question Answering. pp. 73–81 (2021)
84. Tan, W., Zhang, W., Xu, X., Xia, H., Ding, Z., Li, B., Zhou, B., Yue, J., Jiang, J., Li, Y., An, R., Qin, M., Zong, C., Zheng, L., Wu, Y., Chai, X., Bi, Y., Xie, T., Gu, P., Li, X., Zhang, C., Tian, L., Wang, C., Wang, X., Karlsson, B.F., An, B., Yan, S., Lu, Z.: Cradle: Empowering foundation agents towards general computer control (2024), https://arxiv.org/abs/2403.03186
85. Team, O.E.L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., Czarnecki, W.M.: Open-ended learning leads to generally capable agents (2021), https://arxiv.org/abs/2107.12808
86. Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al.: Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032 (2024)
87. Ullman, T.: Large language models fail on trivial alterations to theory-of-mind tasks (2023)
88. Uther, W.: Markov decision processes (2010)
89. Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A.S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., Quan, J., Gaffney, S., Petersen, S., Simonyan, K., Schaul, T., van Hasselt, H., Silver, D., Lillicrap, T., Calderone, K., Keet, P., Brunasso, A., Lawrence, D., Ekermo, A., Repp, J., Tsing, R.: Starcraft ii: A new challenge for reinforcement learning (2017), https://arxiv.org/abs/1708.04782
90. Wade, M., Prime, H., Jenkins, J.M., Yeates, K.O., Williams, T., Lee, K.: On the relation between theory of mind and executive functioning: A developmental cognitive neuroscience perspective. Psychonomic bulletin & review **25**, 2119–2140 (2018)
91. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application (2024), https://arxiv.org/abs/2302.00487
92. Wellman, H.M., Lagattuta, K.H.: Theory of mind for learning and teaching: The nature and role of explanation. Cognitive development **19**(4), 479–497 (2004)
93. Wellman, H.M., Liu, D.: Scaling of theory-of-mind tasks. Child Development **75**(2), 523–541 (2004). https://doi.org/https://doi.org/10.1111/j.1467-8624.2004.00691.x
94. Wilensky, U.: Netlogo itself (1999), http://ccl.northwestern.edu/netlogo/
95. Williams, J., Fiore, S.M., Jentsch, F.: Supporting artificial social intelligence with theory of mind. Frontiers in Artificial Intelligence **5** (2022). https://doi.org/10.3389/frai.2022.750763
96. Wimmer, H., Perner, J.: Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983). https://doi.org/https://doi.org/10.1016/0010-0277(83)90004-5
97. Xu, H., Zhao, R., Zhu, L., Du, J., He, Y.: Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models (2024), https://arxiv.org/abs/2402.06044
98. Zhou, P., Madaan, A., Potharaju, S.P., Gupta, A., McKee, K.R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., Upadhyay, S., Faruqui, M.: How far are large language models from agents with theory-of-mind? (2023), https://arxiv.org/abs/2310.03051

99. Zhu, Y., VanderHoeven, H., Lai, K., Bradford, M., Tam, C., Khebour, I., Brutti, R., Krishnaswamy, N., Pustejovsky, J.: Modeling theory of mind in multimodal hci. In: Kurosu, M., Hashizume, A. (eds.) Human-Computer Interaction. pp. 205–225. Springer Nature Switzerland, Cham (2024)