TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 00 (2021) 1–22
© 2021 Cognitive Science Society LLC
ISSN: 1756-8765 online
DOI: 10.1111/tops.12553

This article is part of the topic “Cognition-Inspired Artificial Intelligence,” Daniel N. Cassenti, Vladislav D. Veksler and Frank E. Ritter (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview.

Theory of Mind From Observation in Cognitive Models and Humans

Thuy Ngoc Nguyen, Cleotilde Gonzalez

Dynamic Decision Making Laboratory, Social and Decision Sciences Department, Carnegie Mellon University

Received 26 November 2020; received in revised form 31 May 2021; accepted 1 June 2021

Abstract

A major challenge for research in artificial intelligence is to develop systems that can infer the goals, beliefs, and intentions of others (i.e., systems that have theory of mind, ToM). In this research, we propose a cognitive ToM framework that uses a well-known theory of decisions from experience to construct a computational representation of ToM. Instance-based learning theory (IBLT) is used to construct a cognitive model that generates ToM from the observation of other agents’ behavior. The IBL model of the observer distinguishes itself from previous models of ToM that make unreasonable assumptions about human cognition, are hand-crafted for particular settings, complex, or unable to explain a cognitive development of ToM compared to human’s ToM. The IBL model learns from the observation of goal-directed agents’ behavior in a gridworld navigation task, and it infers and predicts the behaviors of the agents in new gridworlds across different degrees of decision complexity in similar ways to the way human observers do. We provide evidence for the alignment of the IBL observer’s predictions under various levels of decision complexity. We also advance the demonstration of the IBL predictions using a classic test of false beliefs (the Sally–Anne test), which is commonly used to test ToM in humans. We discuss our results and the potential of the IBL observer model to improve human–machine interactions.

Keywords: Cognitive model; Machine theory of mind; Instance-based learning theory; False beliefs; Human experiments

Note: An initial version of the model presented here appeared at the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020) (Nguyen & Gonzalez, 2020a).

Correspondence should be sent to Cleotilde Gonzalez, Social and Decision Sciences Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: coty@cmu.edu

1. Introduction

Theory of mind (ToM) refers to the ability to infer and interpret the beliefs, desires, and intentions of others (Premack & Woodruff, 1978; Rusch, Steixner-Kumar, Doshi, Spezio, & Gläscher, 2020). ToM is known to be an essential component of human learning and social cognition, including the acquisition of social norms and social beliefs (MacLean, 2016). For example, social reasoning often relies on the mental models we have of other agents, and this ability develops in humans at a very early age and evolves in more sophisticated ways as the brain matures (Keysar, Lin, & Barr, 2003; Wimmer & Perner, 1983). In this research, we address the challenge of creating computational models of ToM and how do these models emulate human ToM.

Creating computational representations of ToM is a problem that is addressed recently in artificial intelligence (AI) research. Two prominent approaches of computational frameworks of ToM are the Bayesian ToM (BToM) framework (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2011) and the deep learning machine ToM (MToM) (Rabinowitz et al., 2018). Both of these approaches (explained below) rely on *observational* tasks, where computational agents *observe* other agent's actions while they navigate in a gridworld environment. An *acting agent* makes decisions in a gridworld environment, moving around the gridworld in search for a target. The information about the acting agent's actions becomes available to the observer dynamically, from the observation of the acting agent's actions. That is, the observer does not have knowledge of the agent's preferences and intentions a priori, and these must be learned from observation of the actions the agent takes. To make predictions about the agent's behavior, the observer keeps a dynamic update of the agent's beliefs. After observation, the models are queried about the beliefs or asked to predict the next actions of the acting agent.

In this paper, we adopt a similar approach to the Bayesian and deep learning processes, in which ToM is built dynamically from the observation of other agent's actions in a gridworld task. However, rather than using a Bayesian or deep learning model of the observer, we build a cognitive model of the observer using a cognitive ToM framework (*CogToM*). In CogToM, the cognitive model of the observer is based on an existing cognitive theory of decisions from experience, instance-based learning theory (IBLT; Gonzalez, Lerch, & Lebiere 2003). Our goal is to generate an accurate representation of *human* ToM. To accomplish that goal, we use a model that relies on cognitive mechanisms that can represent human beliefs and human strategic reasoning about another agent's beliefs. An IBL model of the observer generates predictions about the acting agent's behavior and the agent's next action. We evaluate the predictions that the IBL model makes under various conditions of decision complexity and the amount of information regarding the acting agent's actions, using human data collected in an experiment in which humans act as observers. We also advance the IBL model to be able to predict false beliefs of the acting agent, an ability that is essential of ToM (Baron-Cohen, Leslie, & Frith, 1985).

1.1. Bayesian and deep learning ToM frameworks

In the BToM framework, Baker et al. (2017) formalize the agent's behavior in a gridworld environment as a partially observable Markov decision process (POMDP) while considering the classical rational view of maximizing the expected utility. The gridworld is a sequential decision-making task where agents navigate through a grid to search for targets. The BToM is a generative probabilistic model that indicates how observable actions are attributed to the acting agent's mental states, for example, beliefs and/or desires. Then, by applying Bayes' rule, these models are inverted to infer and predict the agent's mental state from the observation of their actions, assuming an exact observation and probability calculation.

The MToM architecture proposed by Rabinowitz et al. (2018) emphasizes the use of a deep learning model and the scalability of the computational approach. In this work, an observer called ToMnet is represented by a deep neural network consisting of three nets. The "character net" parses the episodes of many agents in many environments to learn about the common behavior of all agents in the training set, which results in a "character embedding." Then the character embedding is combined with "agent embedding" generated from the observations of a single agent at the test time to represent the "mental state net" of that agent. These two embeddings become the inputs of the "prediction net" that is then used to infer and make predictions regarding the agent's future behavior. The authors offer a series of simulation experiments with the observer's predictions, including the classic test of recognition of false beliefs, the Sally–Anne test (Baron-Cohen et al., 1985). The Sally–Anne test is a prominent way to assess ToM in models and human participants (Baker et al., 2017; Wimmer & Perner, 1983). The simulation results demonstrate the development of ToM in the ToMnet observer, but these results are not validated against human observers.

Our proposed approach, the CogToM, is similar to the BToM and MToM frameworks in two respects: (1) it relies on the observation of an agent acting in a gridworld environment and (2) it proposes a dynamic model that develops ToM from the observation of the other agent's actions. However, generally, models that assume accurate observations or base their calculations on the assumption of utility maximization tend to deviate from actual human behavior in many decision-making tasks (Kahneman, Slovic, Slovic, & Tversky, 1982). Instead, humans are often boundedly rational (Simon, 1956), and their decisions are constrained within human cognitive capabilities for storing and retrieving information from memory (Gonzalez et al., 2003). CogToM provides a *cognitive* model of the observer, rather than a Bayesian or deep learning model, and we demonstrate that CogToM provides predictions that are in agreement with human observers.

1.2. Layout of steps forward

We first present the proposed CogToM, IBLT, and the IBL model of the observer in a gridworld task, and then we present two experiments. In Experiment 1, we demonstrate that the IBL observer is able to infer and make predictions about the actions of reinforcement learning (RL) agents acting in a gridworld under different degrees of decision complexity and observation of the agent's actions. Next, we present a human experiment designed to

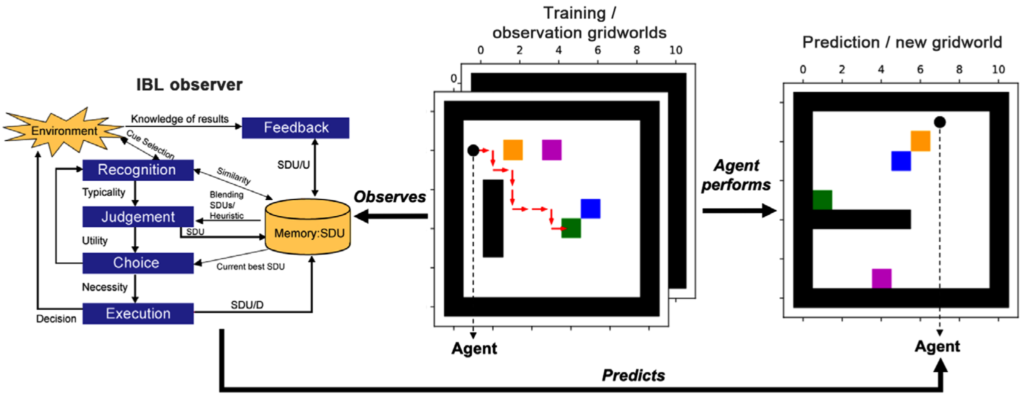


Fig. 1. CogToM framework.

corroborate the predictions from the IBL observer model. We provide experimental evidence that the IBL observer can make inferences that align very well with a human observer.

Experiment 2 is designed to advance CogToM by testing the IBL model's ability to predict *false beliefs* of acting agents of various types. To test false beliefs, we design a test analogous to the Sally–Anne task in the gridworld (Baker et al., 2017; Wimmer & Perner, 1983), where the preferences of acting agents (random, RL agents and IBL) are tricked under conditions of visibility or not, and the IBL observer is asked to make predictions about the behavior of the acting agents of different types.

The two experiments suggest that the IBL observer can generate sensible ToM from the observation under different levels of complexity and observability of the agent's actions in the gridworld.

2. CogToM: A cognitive machine theory of mind framework

In the cognitive machine theory of mind (CogToM) framework (Nguyen & Gonzalez, 2020a) (see Fig. 1), the *observer* is a cognitive model that builds ToM dynamically, by observing (fully or partially) the actions that an *agent* takes in a gridworld. The observer makes predictions regarding the agent's future behavior, such as a next-step action or the agent's goal in a new gridworld.

2.1. A gridworld task

A gridworld is a sequential decision-making problem wherein agents move through a $N \times N$ grid to search for targets and avoid obstacles (11×11 grid following (Rabinowitz et al., 2018)). It contains randomly located obstacles (black bars), and the number of obstacles varies from zero (no obstacles) to six with the size of 1×1. In each grid, there are four goals of different values, represented by four colored objects (blue, green, orange, and purple), which are put at non-overlapping locations. Starting at a random position (i.e., (x, y)), the agent (black dot) makes sequential decisions about which actions to take (i.e., up, down, left, right) to reach one of the four objects. A sequence of moves from the initial location to the end location forms a *trajectory* (dotted red line), which is produced by the strategy (the sequence of decisions) that the agent takes.

Mathematically, a gridworld can be formulated as a POMDP as in Rabinowitz et al. (2018). Each POMDP $\mathcal{M}_j$ has a state space S_j, and each square in the grid is called a state $s \in S_j$. At each state s, an agent $\mathcal{A}$ is required to take an action a from an action space A_j. Each agent follows their policy (i.e., strategy), to decide how to move around the grid. By executing its policy π_k in the gridworld $\mathcal{M}_j$, the agent $\mathcal{A}$ creates a trajectory denoted by $\mathcal{T}_j = \{(s_t, a_t)\}_{t=0}^{T}$, where T is the number of steps taken by the agent. If the agent has a full observation of the grid, POMDP is referred to as Markov Decision Process (MDP).

2.2. IBLT and the IBL observer model

IBLT (Gonzalez et al., 2003) provides a decision-making algorithm and a set of cognitive mechanisms used to implement computational models. In essence, IBL models make decisions by generalizing the outcomes of past experiences (i.e., instances) according to their similarity to a current decision situation. An "instance" is a memory unit that results from the potential alternatives evaluated. These are memory representations consisting of three elements: a situation (a set of attributes that give a context to the decision, or state s); a decision (the action taken corresponding to an alternative in state s, or action a); and a utility (expected utility or experienced outcome x of the action taken in a state). IBLT leverages on some of the cognitive mechanisms of a well-known cognitive architecture, ACT-R (Anderson & Lebiere, 2014), to calculate the way declarative memory elements are retrieved, activated, and used. Generally, the memory mechanisms in ACT-R have been robustly validated against human data, and IBL models have successfully accounted for the inferences and choices made by humans in a wide variety of decision tasks (Gonzalez, 2013; Gonzalez & Dutt, 2011; Gonzalez et al., 2003; Hertwig, 2015; Lejarraga, Dutt, & Gonzalez, 2012).

In the gridworld task, an *instance* is a triplet (s, a, x), where x is the observed or expected outcome resulting from taking action a at state s (i.e., the state is the location of the agent, defined by the x–y coordinates) in a grid. An option $k = (s, a)$ is defined by the action a taken in state s. At time t, there may be $n_{k,t}$ different generated instances $(k, x_{i,k,t})$ for $i = 1, \ldots, n_{k,t}$, corresponding to selecting k and achieving outcome $x_{i,k,t}$. IBL models rely on three main mechanisms: Activation (Eq. 1), probability of retrieval (Eq. 2), and blending (Eq. 3) to make decisions in each time t.

Each instance i in memory has an *activation* value, which represents how readily available that information is in memory according to several factors including similarity and recency (Anderson & Lebiere, 2014). In the IBL model of the gridworld task, we consider a simplified version of the ACT-R's activation equation which only captures how recently and frequently instances are activated:

$$A_{i,k,t} = \ln\left(\sum_{t' \in T_{i,k,t}} (t - t')^{-d}\right) + \sigma \ln \frac{1 - \xi_{i,k,t}}{\xi_{i,k,t}}. \tag{1}$$

In Eq. 1, d and σ are the decay and noise parameters, respectively, and $T_{i,k,t} \subset \{0, \ldots, t-1\}$ is the set of the previous timestamps in which, the instance, i was observed. The rightmost term represents the Gaussian noise for capturing individual variation in activation, and $\xi_{i,k,t}$ is a random number drawn from a uniform distribution $U(0, 1)$ at each timestep and for each instance and option.

Importantly, in all simulations presented in this paper, we used the "default" parameter values commonly used in the literature (noise $\sigma = 0.25$ and decay $d = 0.5$); the parameters were not optimized, and the results are pure predictions from the theoretical principles of IBLT.

Activation of an instance i is used to determine the probability of retrieval of an instance from memory. The probability of an instance i is a function of its activation $A_{i,k,t}$ relative to the activation of all instances:

$$p_{i,k,t} = \frac{e^{A_{i,k,t}/\tau}}{\sum_{j=1}^{n_{k,t}} e^{A_{j,k,t}/\tau}}, \tag{2}$$

where τ is the Boltzmann constant (i.e., the "temperature") in the Boltzmann distribution. For simplicity, τ is often defined in IBL choice models as a function of the same σ used in the activation equation $\tau = \sigma\sqrt{2}$ as in Lejarraga et al. (2012).

The expected utility of the option k is calculated based on a mechanism called *blending* (Lebiere, 1999), which in IBL choice models is specified as an expected value, using the probability of memory retrieval (Eq. 2) and the outcomes stored in each instance (Gonzalez & Dutt, 2011; Gonzalez, Lerch, & Lebiere, 2003; Lejarraga et al., 2012):

$$V_{k,t} = \sum_{i=1}^{n_{k,t}} p_{i,k,t} x_{i,k,t}. \tag{3}$$

The IBL choice rule is to select the option with the highest blended value.

The IBL observer learns from the observation of full or partial information of the decisions made by the acting agent. The "past experience" of the IBL observer is implemented by inserting "pre-populated instances" in the model's memory, a mechanism that has been used to emulate initial expectations in a choice task (Lejarraga et al., 2012). More precisely, each observed trajectory $\mathcal{T}_j$ produced by an agent $\mathcal{A}$ following its policy π in the gridworld $\mathcal{M}_j$ is inserted in the IBL observer's memory. Presumably, each agent has their true reward signal R that defines their desired goal, reflected in the path taken in the task. Derived from the observable actions of the agent, the observer first needs to infer the agent's true reward function which is inaccessible to the IBL observer. Then, based on the inferred reward, the IBL observer makes the prediction about the agent's behavior in the new environment. Simply put, the goal of the observer is not only to infer the agent's objectives or rewards but also to learn the path the agent would take in a new environment. This differentiates our work from the approach of inverse reinforcement learning (Ng et al., 2000), which is merely aimed at finding a reward function that explains the given agent's history of behavior.

2.3. Computational models of acting agents

We consider three different types of acting agents that play in the gridworld: *random*, RL, and IBL agents. The random agent serves as a baseline to compare the IBL observer's predictions with the predictions made by the IBL observer regarding the other learning agents (RL or IBL).

Random agent. A random agent $\mathcal{A}$ selects an action a in state s based on the probability $\pi(a|s)$. More precisely, the policy of $\mathcal{A}$ is drawn from a Dirichlet distribution $\pi \sim Dir(\alpha)$ with concentration parameter α, wherein $\sum_{a \in A} \pi(a|s) = 1$ and $\pi(a|s) > 0$.

Reinforcement learning agent. A RL agent adopts a tabular form of *Q-learning* algorithm, a quintessential temporal difference approach (Sutton et al., 1998). In general, the goal of the RL agent $\mathcal{A}$ is to estimate the optimal state-action values referred to as Q-values, where $Q(s, a)$ returns the expected future reward of action a at state s. Initially, all Q-values are set to zero and then are iteratively updated. Given enough iterations, the agent can learn the optimal Q-values denoted by $Q^*(s, a)$, and for each state s the agent selects the action having the highest Q-value, $\pi^*(s) = \text{argmax}_a Q^*(s, a)$. As our main concern is in the performance of the IBL observer rather than the agents, we only explore *Q-learning* agents since the temporal difference method corresponds closely to the learning behaviors of humans (Sutton et al., 1998), though the implementation of different RL algorithms is entirely possible.

IBL agent. An IBL agent acting in the gridworld uses the memory and learning mechanisms described above. The IBL agent first makes a calculation of the expected utility of each potential action a the agent $\mathcal{A}$ will take at state s, using the *blended* value (Eq. 3), then it selects the action with the highest blended value. Importantly, the agent only gets the outcome at the end of each episode when a goal is reached, typically entailing a sequence of steps. Thus, the IBL agents must update the utility of the instances generated at each step before reaching a goal. To that end, we employ a mechanism of credit assignment in the IBL model, where the actual observed outcome is assigned equally to all actions taken in a trajectory. That is, considering the trajectory $\mathcal{T} = \{(s_t, a_t)\}_{t=0}^{T}$ if the $\mathcal{A}$ gets the outcome x' at the end of the episode $(t = T)$ then outcome of executing $\{(s_t, a_t)\}_{t=0}^{T-1}$ is all updated to x', excluding those that result in an agent walking into a wall or obstacle. This is an exploratory mechanism to address the well-known *credit assignment problem* in AI (Minsky, 1961). The development and comparison of credit assignment algorithms in the IBL model are investigated separately (Nguyen, McDonald, & Gonzalez, 2021).

3. Experiment 1: Predicting next step and preferred goal

In Experiment 1, we use a *goal-directed* task in which an IBL observer predicts the actions of an acting agent, a RL model, in a new gridworld after observing the agent's behavior in other gridworld configurations. We first present a simulation experiment involving two main factors: the decision complexity to which the RL agent is subject, and the level of information that the IBL observer has about the RL agent's trajectory.

Next, we validate the predictions of the IBL observer by comparing the results with human data collected in an experiment in which human participants played the role of an observer.

3.1. Simulation experiment

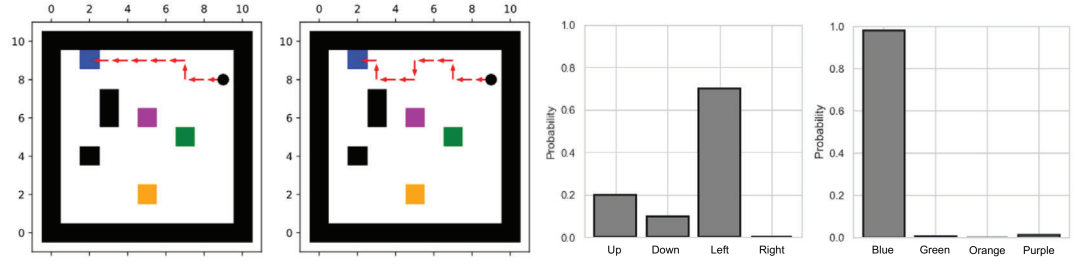
The task of the RL agent in the gridworld was to reach the object (out of four available ones) with the highest value within 31 steps. The RL agent is penalized for each move and for walking into a wall, 0.01 and 0.05, respectively. Consuming any object (even if it is not the highest reward object) leads to the termination of the episode, even if the 31-step limit has not been reached.

More formally, each RL agent $\mathcal{A}$ is driven by a fixed reward function that rewards the agent with $r_o \in (0, 1)$ for consuming an object o. The vector r is drawn from a Dirichlet distribution with concentration parameter $\alpha = 0.01$, whose elements are non-negative and sum to 1. This means only one object will have a high reward, while the others have very small rewards. The discount factor γ and learning rate in the RL agent's model are set to 0.95 and 0.1, respectively.

In turn, given the observation of the RL agent's trajectory in a training gridworld, the IBL observer is tasked with predicting the agent's next action, and the object the agent would consume at the end of the episode in a new gridworld. It is important to stress that even though the RL agent's behavior was observable, its reward function along with its policy were unknown to the IBL observer. Therefore, the IBL observer's mission is to learn to infer which object the RL agent desires to consume, which is determined by its reward function and transferable to a new gridworld. Based on the inference, the IBL observer makes behavioral predictions of the agent's actions in the new gridworld environment.

Before we present the experimental conditions of Experiment 1, we ran a simulation as an example of the results of the IBL observer. This example illustrates a true trajectory of the RL agent in the new gridworld after training (Fig. 2a). The RL agent learns to consume the highest value goal (e.g., the blue target). Fig. 2b shows the predicted trajectory from the IBL observer. The predicted trajectory is less direct than the true trajectory, but ultimately, the IBL observer predicts a path to the highest value goal. Fig. 2c shows the IBL observer's accuracy of the prediction about the first step that the RL agent would take from its initial position (aka. next step action). The model accurately predicts that the RL agent will most likely go "left" in this example. Finally, Fig. 2d presents the IBL observer's prediction of the object that the RL agent will consume at the end of the episode. The IBL observer makes a clear prediction that the RL agent will consume the blue object (the highest value object).

Effects of decision complexity and partial observation of the trajectory. In this experiment, we tested the effects of decision complexity and the information available to the IBL observer regarding the decisions of the RL agent. Decision complexity is inspired by the cost-benefit trade-offs of choices in a gridworld task (Nguyen & Gonzalez, 2020b). The level of complexity is characterized by the difference between the distance from an agent's spawn location to the highest reward object (d) and to the nearest distractor (low-reward object) (d'). The difference is then denoted by $\Delta_d = d - d'$. The larger the value of Δ_d, the more complex the decision is, provided that the high value of Δ_d indicates a tension between consuming the



(a) True trajectory (b) Pred. trajectory (c) Pred. next-step (d) Pred. consumption

Fig. 2. (a) Illustrates the true trajectory of an example RL agent. (b) Illustrates the IBL observer's predicted trajectory of the RL agent given full observations. (c) The results from the IBL observer's prediction of the RL agent's next action in the new gridworld: the probability of taking the action "left" is about 0.7. (d) The results from the IBL observer's prediction for object to consume at the end of the episode in the new gridworld: the probability of consuming the "blue" object is about 0.98.

highest reward object with the longer distance d or going for the distractor with the shorter distance d'.

Here, we considered three levels of complexity: "simple" ($\Delta_d \leq 1$), "complex" ($\Delta_d \geq 4$), and "random" (the value of Δ_d was not controlled, that is the positions of the agent and the objects were generated randomly).

We further inspected the case when the IBL observer was provided with an RL agent's full trajectory in the past gridworld (i.e., *full information*) and when it was limited to observing a partial trajectory (i.e., a single action pair in the past trajectory, *partial information*). For the analysis of partial trajectories, we manipulated the number of past gridworlds ($N_{\text{past}} = 0 \ldots 10$) from which the IBL observer could learn about the RL agent's preferred object.

We experimented with 100 RL agents in each setting, and in each gridworld, the acting agent had 500 training episodes. Likewise, the IBL observer was allowed to have 500 episodes to learn about the acting agent's behavior from the observation before making the predictions in a new gridworld.

Dependent variables . For each level of decision complexity, we assessed the IBL observer's accuracy in a new gridworld after having received full or partial information. More specifically, given the initial location of an agent in a new gridworld, the IBL observer was queried about: (1) the first action that the RL agent would take from its starting position (i.e., next action prediction) and (2) the object the RL agent would consume by the end of the episode (i.e., goal consumption prediction).

The accuracy was measured by the proportion of the IBL observer's accurate predictions relative to the RL agent's true behavior, as illustrated in Fig. 2. Providing that the RL agent using the *Q-learning* algorithm only converges to optimal values under specific conditions (Dayan & Watkins, 1992), the true behavior is defined by the object that the RL agent consumes by following its suboptimal policy with respect to the predefined reward function R. Arguably, the RL agent is not always successful in obtaining the highest reward goal, although it was trained to be competent at the task after a long number of episodes (e.g., 500).

IBL observer training . As the agent's reward function was concealed from the observer, only the observable actions were used for training the IBL observer.

In the full trajectory case, the IBL observer acts according to the probability distribution over the objects consumed by the agent in the past gridworld, to learn about the agent's reward function.

With partial information, the IBL observer was trained to identify the preferred object based on the movement direction of the agent's action. Technically, the past gridworld was divided into four areas by joining the RL agent's location to the four edges of the grid. The area that the agent visited most frequently over the 500 learning episodes was assumed to potentially contain its preferred object. If there was no object in that area, then the preferred object was chosen randomly. Conversely, if it contained more than one object, the preferred object was selected based on its frequency in the target area over N_{past} grids when $N_{\text{past}} > 1$.

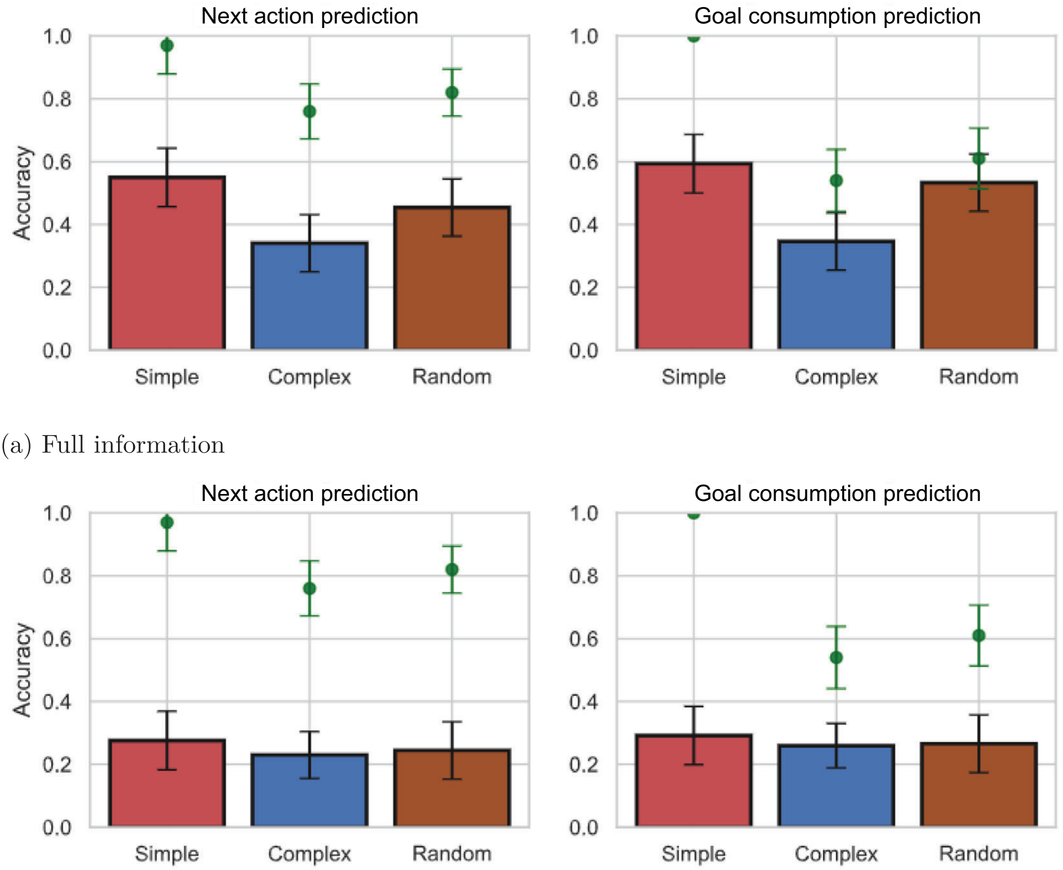
Results. Fig. 3a shows the IBL observer's average accuracy for predicting the RL agents' next action as well as the object consumed in the three levels of decision complexity and under conditions of full and partial information. The green dots represent the RL agents' actual behavior.

In the RL agent's actual behavior, there is an effect of decision complexity on both the next action and the goal consumption, where RL agents are more accurate in simple gridworlds. The IBL observer also predicts the complexity effect, but clearly, the prediction accuracy of the IBL observer is significantly lower than the *actual* behavior of the RL agent. Furthermore, as expected, the IBL observer's prediction accuracy drops given partial observation of information.

More concretely, the average results from the 100 agents show that the mean prediction accuracy in the simple setting is the highest (0.55 ± 0.08 for next action prediction and 0.60 ± 0.09 for goal consumption prediction with 95% confidence level), followed by the random (next action: 0.45 ± 0.09; goal consumption: 0.53 ± 0.09) and complex (next action: 0.34 ± 0.09; goal consumption: 0.35 ± 0.09) conditions.

The complexity effect of the IBL observer's predictions can be explained by the fact that the inference and predictions made by the IBL observer rely on the actions taken by the RL agent. When the environment becomes more challenging, the RL agent is more likely to erroneously consume closer distractors, which leads to the decline in IBL observer's prediction accuracy. When the IBL observer is provided with partial information, the difference in the prediction accuracy of the next action and of the consumed goal across the three levels of complexity is negligible, as shown in Fig. 3b. Given fewer observations, the IBL observer performs considerably worse than with full information.

This experiment demonstrated that the acting agents and the IBL observer are sensitive to the level of decision complexity in the gridworld, and that the IBL observer is additionally sensitive to the amount of information observed from the RL agent. Importantly, given that IBL models are expected to replicate human behavior, the next question is whether human behavior would compare to the IBL observer behavior regarding the RL agent's next actions and goal consumption.



(a) Full information

(b) Partial information $N_{past} = 1$

Fig. 3. Accuracy of the IBL observer's prediction across the three levels of decision complexity, given full and partial information about the RL agents' trajectory. The green dots represent the agents' actual behavior, for comparison against the observer.

3.2. *Human observer experiment*

To evaluate the IBL observer's ToM against human observers, we conducted a behavioral experiment in which humans observed the RL agents in an interactive gridworld task under different levels of decision complexity. In other words, we tested whether human observers' predictions were sensitive to decision complexity and whether they would be comparable to the predictions of the IBL observer.

Experimental setup. A participant of this experiment is an observer tasked with watching the RL agent navigating the gridworld to obtain its preferred object, before being asked to make predictions regarding the RL agent's actions in new gridworlds of various complexity

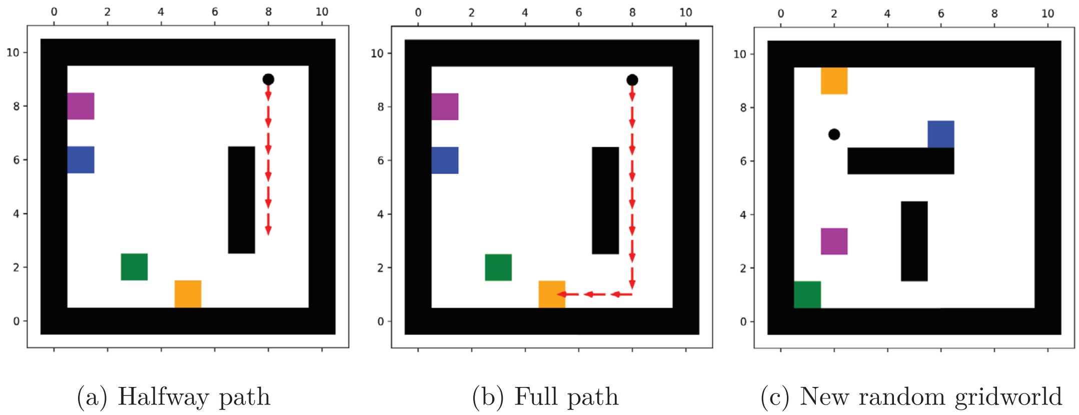


Fig. 4. Examples of gridworld in the observation and test phases. (a) Agent's progress when it is half way through its exploration. (b) Agent's completed progress. (c) New gridworld environment in the test.

levels. Similarly to the simulation experiment, participants were not informed about the RL agent's preferred goal. Rather, participants were asked to observe the agent's to infer and predict the behavior of the RL agent in a new gridworld later.

Participants. All participants were recruited from Amazon Mechanical Turk for 30–45 min in a study preregistered with Open Science Framework.[1] Participants were compensated a base payment of $4.00 and earned up to $1.50 in a bonus payment. After excluding those who failed attention checks, we kept a total of 198 participants (mean ± standard deviation age: 36.11 ± 10.4; 71 females).

Procedure. After providing informed consent and they received instructions, participants executed the task in two phases: (1) an observation (i.e., training) phase and (2) a test phase. Then, they filled out ToM and demographic questionnaires. An attention check (i.e., asking participants to select the left option in a screen with various questions) appeared at random during the task (see the preregistered report for more details on the experimental procedures).

During the training phase, participants were randomly assigned to one of four training sets. Each of the four sets of images was structurally equivalent, and they were only generated to have variability in episodes across participants. Each of these sets includes 20 episodes that were selected from a "random" level of complexity. Each episode involved two animated images. The first image showed the RL agent exploring an 11 × 11 gridworld, and its navigation up to the halfway of its path in that episode. After the first animated image was presented, participants were asked to answer two questions about the image: (a) what direction will the agent go next (i.e., next step: up, right, left, down) and (b) what object the agent will try to consume (i.e., preferred goal). Right after responding to these two questions, participants were shown the animation of the RL agent's completed path in that episode (see Fig. 4a and 4b).

During training, participants had 20 rounds and got a full view of the gridworld, in contrast to the IBL observer that was trained on 500 episodes but can only learn about the environment through the agent's actions revealed one step at a time. Despite the differences during the learning phase, the IBL observer and the human observers can be analyzed for their performance in the test phase that was equivalent for the IBL and human observers.

The test phase appeared immediately after the training phase. Participants were instructed to predict the agent's next step and the object to be consumed in six novel gridworlds, two were simple, two were complex, and two were random (see example in Fig. 4c).

Finally, participants were asked to answer a ToM questionnaire and a demographic (i.e., age, gender, and level of education) questionnaire. The ToM questionnaire included the measurement of three skills relevant to ToM: recursive thinking, judgment of false beliefs, and perspective taking.

The recursive thinking questionnaire involves two stories, each followed by second- and third-order recursive questions inspired by (O'Grady, Kliesch, Smith, & Scott-Phillips, 2015). After reading each story about a social situation (i.e., a story about babysitting and workplace romances), participants answered one second-order (e.g., X believes that Y thinks …) and one third-order recursive thinking question (e.g., X believes that Y thinks that Z believes …).

The false beliefs questionnaire was a measure similar to the Sally–Anne test developed for adults by Liddle and Nettle (2006) and Valle, Massaro, Castelli, and Marchetti (2015). We developed a version of this task involving watching two videos about a social scenario (i.e., a story about playing football and three brothers). After watching each of the videos, participants answered one second-level (e.g., X believes that Y thinks …) and one third-level false-belief question (e.g., X believes that Y thinks that Z believes …).

The perspective-taking questionnaire was designed according to Hegarty and Waller (2004), and it involved four questions; each question presented eight icons (stop sign, car, etc.) positioned in a circle. Participants were asked to imagine a person standing at various places in the grid and then to identify what direction that person would need to head to point to a third object.

Dependent variables. Participants were asked to predict in the new gridworld (a) the action of the RL agent (i.e., up, down, left, right) at the start of the agent's navigation and (b) the object that the agent would consume. We determined the proportion of the correct responses to the initial action and the proportion of correct predicted beliefs (agent's consumed goal) by comparing the participant's responses to the true behavior of the agents in the test phase.

3.2.1. Results

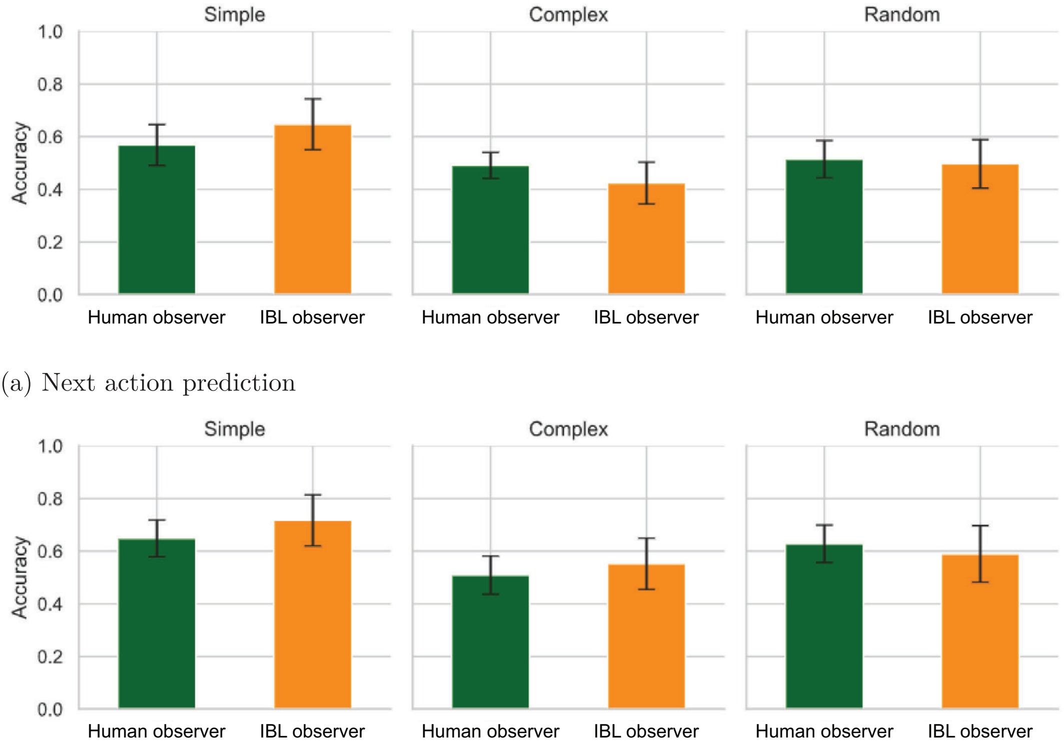
We conducted a one-factorial, repeated measures ANOVA to test for differences in human participants' prediction accuracy across the three levels of complexity during the test phase. The results show that there was a statistically significant difference in the accuracy of next action prediction ($F(2, 394) = 3.63$, $p = .029$) and goal consumption prediction ($F(2, 394) = 10.6$, $p = 5.4e{-}5$) across the three levels of complexity. Table 1 reports the pairwise comparison between each condition, adjusting the p-values using the Bonferroni test.

Just as the IBL observer's results predicted, the human data revealed a significant difference between the simple and complex settings on the next action accuracy. To visualize the IBL observer and the human predictions, Fig. 5 displays the average prediction accuracy of the IBL observer and of the human observers with respect to the next action and goal consumption prediction accuracy in each of the three levels of decision complexity. These averages were

Table 1
Accuracy of next action and goal consumption prediction

	Group1	Group2	n1	n2	*t*-score	*p*	*p*.adj	*p*.adj.signif
Next action	complex	random	198	198	−0.503	.616	1.00	ns
	complex	simple	198	198	−2.79	.006	.017	*
	random	simple	198	198	−1.86	.064	.191	ns
Goal consumption	complex	random	198	198	3.26	.001	.004	**
	complex	simple	198	198	−5.09	.000	.000	****
	random	simple	198	198	−1.05	.294	.882	ns

Notes. ns: not significant; $*$: significant differences .



(a) Next action prediction

(b) Goal consumption prediction

Fig. 5. Comparing IBL observer and mean human ($n = 198$) with respect to next action and goal consumption predictions in three levels of decision complexity.

compared using a non-parametric Wilcoxon signed-rank test. The *p*-values of the tests across the three levels of complexity are all greater than .05, suggesting that there is no statistically significant difference in the predictions of the human observers and those of the IBL observer in terms of next action and goal consumption prediction.

Table 2
An overview of mapping the Sally–Anne test onto the gridworld task

Sally–Anne Test	Gridworld Task
a) Sally places a marble in a basket	a) Agent $\mathcal{A}$ is trained to be a blue-object preferring agent
b) Sally moves away	b) $\mathcal{A}$ is forced to reach a subgoal first
c) Anne puts the marble to a box	c) Location of the preferred blue object is swapped
d) Where will Sally look for her marble when returning (the basket or the box)?	d) At the subgoal position, where will $\mathcal{A}$ go to find the preferred blue object (the original or new location)?

Exploratory analysis of humans' ToM questionnaires. We performed correlations between the human participants' responses to the ToM questionnaires (i.e., recursive thinking, false-belief, and perspective taking) and their accuracy in predicting next step action and goal consumption in the test phase. The mean and standard deviation values for each of the questionnaires are, respectively, recursive thinking: 3.52 ± 0.9, false-belief: 2.74 ± 0.91, and perspective taking: 2.53 ± 1.41. There were a total of four questions for each questionnaire, and one point was given to each correct answer. Hence, the maximum score that the participant can get for the three questionnaires is 12.

The Kendall rank-based correlation test resulted in no significant correlations between any of the prediction accuracy measures and the ToM questionnaire responses.

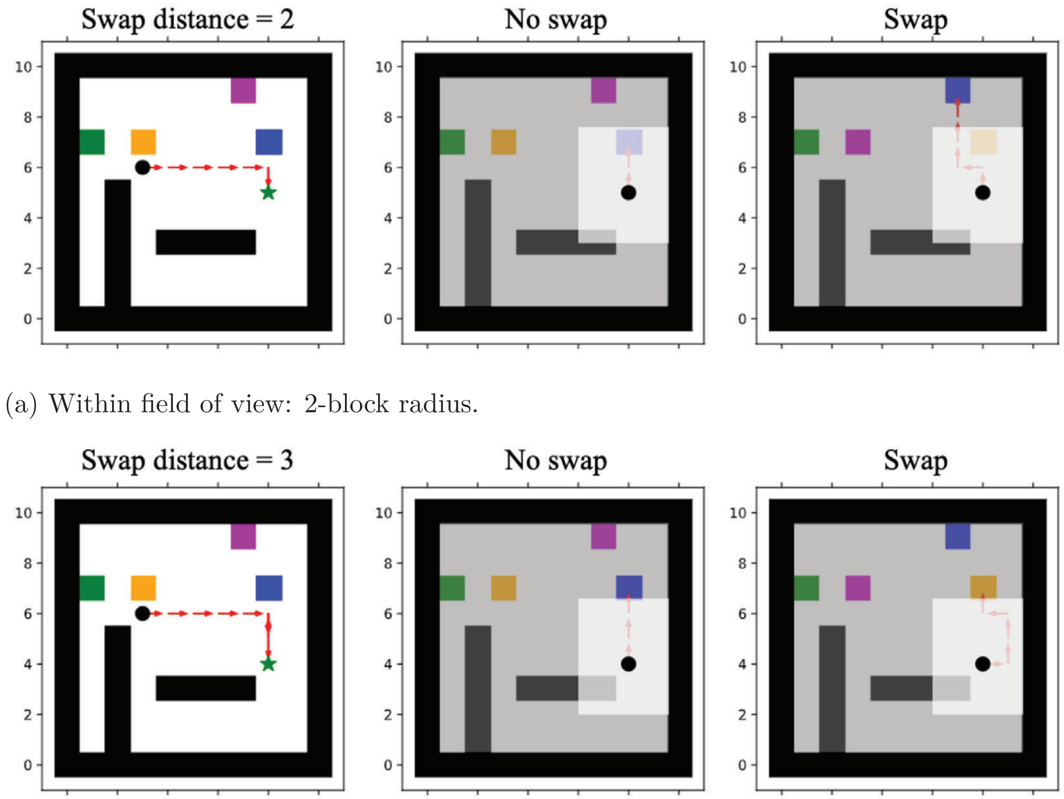
4. Experiment 2: Predicting false-beliefs from observation

Experiment 1 suggests that the IBL observer's predictions are well aligned with human observers' predictions regarding the next action and goal consumption in the gridworld.

In Experiment 2, we advance the IBL observer to predict *false beliefs* of the observed agent. In false belief reasoning, the state of the environment changes without the awareness of the acting agent while the observer is aware of the change. In other words, there is a mismatch between the observer and the acting agent's knowledge of the environment. When an observer is queried about the beliefs of the acting agent, it should be able to distinguish between their own correct belief and the acting agent's *false belief* and respond based on their representation of the acting agent's mental state.

A common test of false beliefs linked to ToM is the Sally–Anne test (Baker et al., 2017; Baron-Cohen et al., 1985; Goodman et al., 2006; Wimmer & Perner, 1983). Inspired by the work of Rabinowitz et al. (2018), we tested the IBL observer for the recognition of agents' false beliefs using an adaptation of the Sally-Anne test to the gridworld task, which is described in Table 2.

Initially, we generated a set of gridworlds in which an agent $\mathcal{A}$ was trained to prefer a blue object. However, the agent is required to reach a subgoal (i.e., an additional object represented by a star in Fig. 6a and 6b) before returning to consume its preferred blue object. During this time, the location of the preferred object was either swapped or not swapped. If the swap event occurred, the locations of the four objects were randomly permuted. Eventually, the



(a) Within field of view: 2-block radius.

(b) Outside field of view: 3-block radius.

Fig. 6. Effect of the swap event on the agent's belief when the event happens within and outside its view. The dark and light gray colors represent the areas outside and within the agent's current field of view.

IBL observer was asked to predict whether or not the agent $\mathcal{A}$ would return to the original location of the blue object in the *no-swap* and *swap* conditions.

As the subject of the test, the IBL observer was aware of the changes in the gridworld (i.e., the swap event); hence, it is expected to indicate that if the agent $\mathcal{A}$ sees the swap, then it will not go back to the original location, but if the agent is not aware of the swap then it will return to the original location of the blue object. This test will signify that the IBL observer is able to model the agent's true and false beliefs.

To determine whether an agent $\mathcal{A}$ sees the swap or not, we introduced a distance variable ($dist$). If the swap event takes place within the agent's view, that is, the distance between the agent's location and the preferred object's location is within a 2-block radius, the agent sees the change, and hence its policy is updated accordingly. Conversely, if the swap event occurs outside the agent's view, the agent's policy remains unchanged, which exhibits a sign of a false belief.

The effect of the swap event on the agent's belief when the event happens within and outside its view is illustrated in Fig. 6. When the swap event is observable (Fig. 6a), it leads to a change in the agent's belief, that is, the agent goes to the new location of the preferred blue object (*true belief*). When the swap event is unobserved (Fig. 6b), it does not lead to a change in the agent's belief, that is, the agent returns to the original location of the preferred blue object without knowing that the blue object has been moved (*false belief*).

Independent variables. In this experiment, we used three kinds of acting agents: random, RL, and IBL agents. We included an IBL model as an acting agent to explore whether the IBL observer would be more accurate in developing ToM of an IBL agent compared to the ToM of a RL and random models. By definition, the RL and random agents are less aligned with the IBL observer's beliefs. Thus, we expected that the IBL observer would make more accurate predictions regarding the false beliefs of the IBL agent compared to the predictions the IBL observer would make for the RL and random agents.

The experiment was run on 100 different agents corresponding to 100 gridworlds of each type. Both the acting agent and the IBL observer had 500 learning episodes in each gridworld. Additionally, we examined the effect of the swap event on the acting agents' behavior and on the IBL observer's predicted behavior by varying the swap distance: $dist = 1 \dots 5$.

Dependent variables . To evaluate the impact of the swap event on the agent's behavior as well as on the IBL observer's prediction, we compared how the acting agents behaved in the absence and presence of the swap event and how the IBL observer performed when observing each of the three types of agents. More concretely, we measured the average Jensen–Shannon divergence (D_{JS}) between the probability distribution over the locations associated with the object consumed at the end of the episode in the swap and no-swap conditions. In essence, D_{JS} scores between 0 (i.e., the two distributions are identical) and 1 (i.e., the two distributions are maximally different). Likewise, we measured D_{JS} to study how the swap event would affect the IBL observer's prediction about the acting agents' behavior.

IBL observer training . Since an agent was tasked with consuming the subgoal first and then the preferred object, only the agent's policies in the episodes in which such a condition was satisfied were selected for the IBL observer to learn. Hence, the observer was informed about the distance variable, and it could derive the agent's preferred object from looking at what was consumed after the subgoal. The point here is that the observer must infer the agent's beliefs from just observing how the agent behaved when the swap event occurred and when it did not.

To do that, the observer was trained to observe the relative importance between the swap distance and the ratio of how frequently the agent went back to the preferred object's original location over a certain number of learning episodes (e.g., 500). For instance, if the swap event was observable within the agent's sight, then the agent was less likely to return to the original location (low frequency). In contrast, if the swap event was out of the agent's view, then the frequency of revisiting the original position was high.

4.1. Results

Fig. 7 displays the divergence score $D_{JS}(swap, \neg swap)$ of each of the three types of agents: random, RL, and IBL agents as well as the IBL observer's prediction about the behavior of

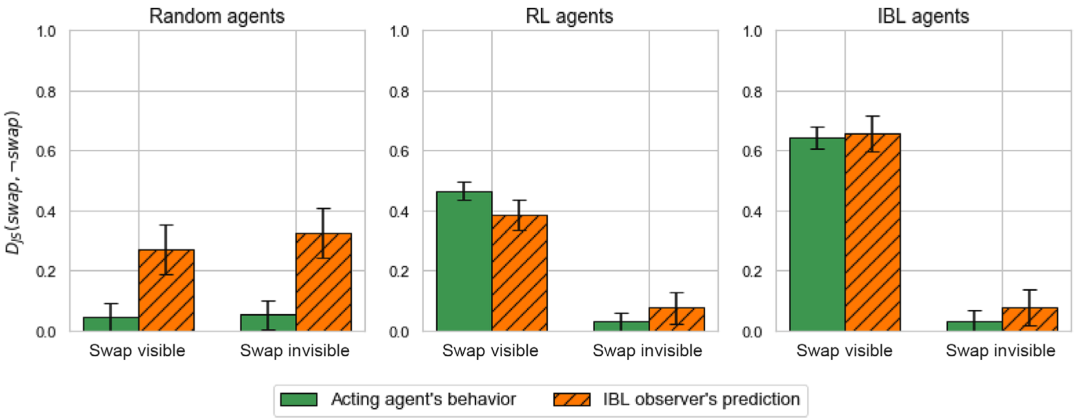


Fig. 7. Effect of swap events with and without visibility factor on the policy of true agents and on the prediction of IBL observer.

each of these agents when the swap is visible and invisible. First, the IBL agent outperforms the RL and random agents in distinguishing the absence and the presence of the swap event when it is visible to the agent ($dis \leq 2$). Concretely, when the swap event occurs within the agents' view, the IBL agent model shows a larger divergence score $D_{JS}(swap, \neg swap)$, given that the probability distribution of the agent's behavior in swap and no swap events is expected to be entirely different. Therefore, the large divergence score $D_{JS}(swap, \neg swap)$ of the IBL agents exhibits better performance compared to RL and random agents. By contrast, when the swap event is invisible to the acting agent ($dist > 2$), none of the three agent types (random, RL, or IBL) is able to recognize the difference between swap and no swap events, leading $D_{JS}(swap, \neg swap)$ to be close to 0. Evidently, the random agents completely fail to differentiate between these two events since its D_{JS} is small and nearly constant regardless the swap distance.

The results obtained from the IBL observer, when observing 100 agents of each type (random, RL, IBL) show that the IBL observer can make predictions that qualitatively resemble the RL and IBL agents' true behaviors. However, the IBL observer was not able to predict the behavior of the random agents correctly due to their random behavior. That is, the random agents were less likely to turn back to the original location when the swap was not visible to the agents, and hence the IBL observer mistakenly learned that the random agents saw the swap event, which results in the high D_{JS} compared to the random agent's actual behavior. This is an interesting and unexpected result which we will turn back to in the discussion.

To quantitatively determine the accuracy of the predictions of the IBL observer on the random, RL, and IBL agents we conducted a two-sample independent t-test to compare the acting agents' behavior and the IBL observer's prediction in terms of D_{js} when the swap event is visible and invisible to the agents. When the swap is visible, there was a significant difference in the performance of random agents and the prediction of IBL observer ($t = 5.9$, $p = 1.3e{-}7 < .05$, and of the RL agents and the prediction of the IBL observer $t = 2.7$, $p = 0.02 < .05$, suggesting that the IBL observer's prediction about the RL agents' behavior was significantly different from the actual behavior of the RL agents. In contrast, we found $t = 2.1$, $p = 0.11 > .05$ regarding the IBL agents, so there was no statistically significant difference in IBL agents' behavior and those predicted by the IBL observer. These results support our hypothesis that the IBL observer can provide better predictions about the IBL agents' behavior than the predictions about the RL and random agents. In the case of the invisible swap event, there was no statistically significant difference in the predictions of IBL observer and the behavior of RL as well as IBL agents (p-values were larger than .05). However, there was strong evidence to conclude that the random agents' performance was significantly different from the IBL observer's prediction ($t = 7.8$, $p = 2.1e{-}7 < .05$).

5. Discussion

We introduced a cognitive model of ToM (CogToM) built upon an existent cognitive theory of decisions from experience, IBLT (Gonzalez et al., 2003). The observer IBL model represents a process by which ToM can develop dynamically, from the observation of other agent's actions in a gridworld task. In contrast to the BToM and MToM architectures (Baker et al., 2017; Rabinowitz et al., 2018), our work is founded on a cognitive theory which aims at capturing human-like assumptions about bounded rationality of a human observer. For example, the IBL model does not assume that a decision is based on the maximization of utility according to Bayes rule in the BToM framework. Instead, IBL models are boundedly rational, by considering decisions constrained within the limitations of human memory (e.g., recency and frequency of information and errors in the retrieval of information).

Also, CogToM is a very simple framework compared to BToM and MToM frameworks. The IBL observer is computationally straightforward, and it does not depend on large amounts of data and complex architectures (Baker et al., 2017; Rabinowitz et al., 2018). Just like human observers, the IBL observer uses simple cognitive mechanisms, learning from observing how the agents explore the environment and base their decisions on frequency, recency, and noise. Furthermore, Experiment 1 showed that the IBL observer's prediction accuracy is sensitive to the level of complexity of the task and to the amount of information observed from the acting agents, similar to how the RL agent acting in the gridworld was sensitive to decision complexity. In complex environments, the acting RL agent has difficulty finding the highest value target as it usually ends up with consuming a nearby distractor. Both the IBL observer and the human observer are sensitive to the decay in accuracy of the RL agent by observation of the agent's actions.

Importantly, the IBL observer is not accurate at predicting the RL agents' performance in the new gridworld, but neither are humans. The corroboration with human behavior revealed that humans are also unable to predict the RL agent's behavior in similar ways as the IBL observer does. In other words, *predicting like humans do* can be a more challenging task than predicting optimally. Just like human observers, the IBL observer's next action and goal consumption accuracy is low; even with full information, the IBL observers' accuracy (and the human observers') is only above chance. Far from being a problem, this result represents an opportunity to determine ways to improve the human observer's accuracy. A step in our future research is to utilize the IBL model to inform optimal models and to determine the ways in which these models can provide advice to the human player.

Experiment 2 advances the kind of predictions that the IBL observer can make. The IBL observer is able to correctly predict the acting agents' false beliefs. Capturing false beliefs is an important aspect in models of ToM (Rabinowitz et al., 2018) and having a cognitive model that emulates human behavior suggests that the IBL observer is able to capture human's false beliefs.

Two interesting observations emerged from Experiment 2. First, the IBL observer can predict the false beliefs of an IBL agent more accurately than it can predict the false beliefs of random and RL agents. When the swap happens within the agent's view, the IBL observer predicts the true beliefs of the IBL agents better than those of RL and random agents. This is because IBL agent performs better than RL and random acting agents in the first place; resulting in higher IBL observer's accuracy when observing IBL acting agents. When the swap occurs outside of the agent's view, the IBL observer's ability to predict the false beliefs of the RL and IBL acting agents is comparable, and it is clearly better than predicting the behavior of Random agents. The reason is that both, IBL and RL agents, fail to distinguish the swap and no swap event when the swap occurs outside their view. The IBL observer has already learned that when the swap event is outside the agent's view, the agent is more likely to go back to the original location (false belief).

Second, the IBL observer was unable to predict the behavior of the random agents. This is because a random agent was less likely to turn back to the original location given a nonvisible swap event, and the IBL observer "thought" that the random agent saw the swap event. This is an interesting and unexpected result given situations in which random agents could be used to deceive humans (Moisan & Gonzalez, 2017).

In summary, we proposed a cognitive approach to creating computational representation of ToM. In contrast to existent approaches in AI research (BToM and MToM), CogToM is founded on human cognition and it has the potential to replicate human ToM accurately because it does not make unrealistic assumptions about human optimal behavior and rationality. Our cognitive approach replicates human ToM, which may be more challenging than generating optimal models of ToM. Humans are known to be boundedly rational (Simon, 1956) and act according to a number of decision biases and deviations from rationality (Gonzalez et al., 2003; Kahneman et al., 1982). The CogToM approach is, therefore, a potentially useful tool to guide decision support and automation human–machine systems.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency and the Army Research Office and was accomplished under Grant Number W911NF-20-1-0006. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the Army Research Office, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein. We thank Jeffrey Flagg for help with the design of the ToM questionnaire, and help with data collection. We also thank other members of the dynamic decision making laboratory for their comments.

Note

[1] https://osf.io/94pyf/?view_only=78ca1f8ddd4848dfa4a45e6f87c34673

References

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. New York, NY: Psychology Press.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*, 2469–2474.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Dayan, P., & Watkins, C. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292.

Gonzalez, C. (2013). The boundaries of instance-based learning theory for explaining decisions from experience. In S. Bestmann (Ed.). *Progress in brain research* (Vol. *222*, pp. 73–98). Amsterdam, The Netherlands: Elsevier.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review*, *118*(4), 523–551.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the twenty-eighth annual conference of the Cognitive Science Society* (Vol. *6*, pp. 1382–1387). Vancouver, Canada: Cognitive Science Society.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*(2), 175–191.

Hertwig, R. (2015). Decisions from experience. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. *1*, pp. 240–267). Chichester, UK: Wiley-Blackwell.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Lebiere, C. (1999). Blending: An ACT-R mechanism for aggregate retrievals. Paper presented at *Proceedings of the Sixth Annual ACT-R Workshop*, Fairfax, VA.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, *4*(3-4), 231–244.

MacLean, E. L. (2016). Unraveling the evolution of uniquely human cognition. *Proceedings of the National Academy of Sciences*, *113*(23), 6348–6354.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8–30.

Moisan, F., & Gonzalez, C. (2017). Security under uncertainty: adaptive attackers are more challenging to human defenders than random attackers. *Frontiers in Psychology*, *8*, 982.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *ICML 2000: Proceedings of the seventeenth international conference on machine learning* (Vol. *1*, p. 2.). New York, NY: ACM Press.

Nguyen, T. N., & Gonzalez, C. (2020a). Cognitive machine theory of mind. In *Proceedings of the 42nd annual meeting of the Cognitive Science Society (CogSci 2020)* (Vol. *1*, pp. 2560–2566). Virtual meeting: Cognitive Science Society.

Nguyen, T. N., & Gonzalez, C. (2020b). Effects of decision complexity in goal-seeking gridworlds: A comparison of instance-based learning and reinforcement learning agents. In T. C. Stewart (Ed.), *Proceedings of the 18th international conference on cognitive modelling (ICCM 2020)* (pp. 174–179). University Park, PA: Applied Cognitive Science Lab, Penn State.

Nguyen, T. N., McDonald, C., & Gonzalez, C. (2021). Credit assignment in humans and machines: How a model can learn like humans and act optimally. Technical report, Carnegie Mellon University.

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, *36*(4), 313–322.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S., and Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.

Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., & Gläscher, J. (2020). Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, *146*, 107488.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Sutton, R. S. & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. *135*). Cambridge: MIT Press.

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of mind development in adolescence and early adulthood: the growing complexity of recursive thinking ability. *Europe's Journal of Psychology*, *11*(1), 112.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab of this article:

Supplementary Material