Cognitive Machine Theory of Mind

Thuy Ngoc Nguyen (ngocnt@cmu.edu)
Dynamic Decision Making Laboratory, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213 USA

Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213 USA

Abstract

A major challenge for research in Artificial Intelligence (AI) is to develop systems that can infer humans' goals and beliefs when observing their behavior alone (i.e., systems that have Theory of Mind, ToM). In this research we use a theoretically-grounded, pre-existent cognitive model to demonstrate the development of ToM from observation of other agents' behavior. The cognitive model relies on Instance-Based Learning Theory (IBLT) of experiential decision making, that distinguishes it from previous models that are hand-crafted for particular settings, complex, or unable to explain a cognitive development of ToM. An IBL model was designed to be an observer of agents' navigation in gridworld environments and was queried afterwards to predict the actions of new agents in new (not experienced before) gridworlds. The IBL observer can infer and predict potential behaviors from just a few samples of agents' past behavior of random and goal-directed reinforcement learning agents. Furthermore the IBL observer is able to infer the agent's *false belief* and pass a classic ToM test commonly used in humans. We discuss the advantages of using IBLT to develop models of ToM, and the potential to predict human ToM.

Keywords: cognitive model; machine theory of mind; instance-based learning theory.

Introduction

Theory of mind (ToM) refers to the ability of humans to infer and understand the beliefs, desires, and intentions of others (Premack & Woodruff, 1978). ToM is known to develop very early in life (Wimmer & Perner, 1983; Keysar, Lin, & Barr, 2003) and it is one of the most important social skills used to predict others' behavior and intentions, and to theorize about others' beliefs and desires in future situations.

Since its origins, Artificial Intelligence (AI) attempted to "replicate" various human behaviors in computational form, aiming at passing an imitation game (i.e., *Turing Test*) (Lake, Ullman, Tenenbaum, & Gershman, 2017; Turing, 1950): where a machine behavior would be indistinguishable from that of a human. AI work on ToM has investigated how humans "mentalize" robots (machines more generally) and how human ToM develops when interacting with machines rather than other humans (Banks, 2019). While this work is extremely relevant for developing machine representations of ToM, it does not address the major problem of how to build an algorithm that can develop ToM from the limited observation of other agents' actions; a capability that humans excel at (Lake et al., 2017; Botvinick et al., 2017).

Recently, researchers built computational architectures of ToM. A notable example is the work of Baker and colleagues (Baker, Saxe, & Tenenbaum, 2011; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). In this work, researchers developed a Bayesian ToM (BToM) model that is able to predict and attribute beliefs and desires of other agents, given the observation of their actions. The BToM uses Bayes' probabilities and an assumption of utility maximization (i.e., human rationality) to determine the posterior probability of "mental states". Another recent example is the work of Rabinowitz et al. (2018), who developed a Machine ToM (MToM) architecture involving three modules: a *character net* that parses agents' past trajectories of navigation in gridworlds; a *mental state* net, which parses agents' trajectories in recent episodes, that are then used by the *prediction net* which is queried regarding future behaviors of new agents. These authors offer a set of tests of the observer's predictions regarding various types of agents, and a test of recognition of false beliefs, the Sally-Anne test (Wimmer & Perner, 1983).

Our research builds on these efforts, making the following contributions. First, we present a *Cognitive* Machine Theory of Mind (CogToM) framework that relies on a general cognitive theory of decisions from experience, Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, & Lebiere, 2003). Our approach is different from the standard computational models of ToM, summarized above, in that it uses the IBL process and the formulations of the ACT-R architecture (Anderson & Lebiere, 2014) for memory-based inference to demonstrate how ToM develops from observation of other agents' actions. Second, we demonstrate that an IBL model of an observer (i.e, IBL observer) is able to explain the inferences made by three types of acting agents in gridworlds: Random, Reinforcement Learning (RL), and IBL agents. Third, we find that the IBL observer predicts beliefs and actions of IBL acting agents more accurately than it predicts the beliefs and actions of RL or Random agents.

Instance-Based Learning Theory

IBLT is a theory of decisions from experience, developed to explain human learning in dynamic decision environments (Gonzalez et al., 2003). IBLT provides a decision making algorithm and a set of cognitive mechanisms used to implement computational models. The algorithm involves the recognition and retrieval of past experiences (i.e., instances) according to their similarity to a current decision situation.

An "instance" in IBLT is a memory unit, that results from

the potential alternatives evaluated. These are memory representations consisting of three elements: a situation (S) (set of attributes that give a context to the decision, or state s); a decision (D) (the action taken corresponding to an alternative in state s, or action a); and a utility (U) (expected utility u or experienced outcome x of the action taken in a state).

IBLT relies on sub-symbolic mechanisms that have been discussed extensively (e.g. (Gonzalez et al., 2003; Gonzalez & Dutt, 2011; Gonzalez, 2013)), but we summarize here for completeness. Each instance i in memory has a value of *Activation*, which represents how readily available that information is in memory (Anderson & Lebiere, 2014). A simplified version of the Activation equation captures how recently and frequently the considered instances are activated:

$$A_i = \ln\left(\sum_{t' \in \{1..t-1\}} (t-t')^{-d}\right) + \sigma \ln \frac{1-\gamma_i}{\gamma_i}, \qquad (1)$$

where d and σ are respectively the decay and noise parameters; t' refers to the previous timestamp in which the outcome of instance i was observed resulting from choosing an action a at state s. The rightmost term represents the Gaussian noise for capturing individual variation in activation, and γ_i is a random number drawn from a uniform distribution $U(0,1)$.

Activation of an instance i is used to determine its memory retrieval probability:

$$p_i = \frac{e^{A_i/\tau}}{\sum_l e^{A_l/\tau}}, \qquad (2)$$

where $\tau = \sigma\sqrt{2}$ representing the variability in recalling instances from memory, and l refers to the index of all stored instances to normalize p_i.

The expected utility of taking action a at state s is calculated through a mechanism in IBLT called *Blending*:

$$V(s,a) = \sum_{i=1}^{n} p_i x_i. \qquad (3)$$

Essentially, the blended value is the sum of all the outcomes weighted by their probability of retrieval, where x_i is the outcome stored in an instance i associated with taking action a at state s; p_i is the probability of retrieving the instance i from memory; and n is the number of instances containing the different outcomes for taking action a at state s up to the last time.

The choice rule in the model is to select the action a that has the maximum blended value.

CogToM: A Cognitive Machine Theory of Mind Framework

In the Cognitive Machine Theory of Mind (CogToM) (Fig 1), an *observer* is a cognitive model that builds ToM by observing the actions of *agents* that play 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. The observer should be able to accomplish ToM given full or partial observation of the agent's action traces in past gridworlds. The observer model in CogToM is built according to IBLT (Gonzalez et al., 2003).

Gridworld

A gridworld is a sequential decision making problem wherein agents move through a $N \times N$ grid to search for targets and avoid obstacles. We use gridworlds of 11×11 size following (Rabinowitz et al., 2018) (see Fig 1). A gridworld 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 as four colored objects (blue, green, orange, and purple), which are put at random locations that do not overlap the obstacles. Starting at a random position (i.e., (x,y)), the agent (black dot) makes sequential decisions about the 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.

Generally, a gridworld task can be formulated as a Partially Observable Markov Decision Process (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 $\mathcal{A}_k$ 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 $\mathcal{A}_k$ creates a trajectory denoted by $\mathcal{T}_{kj} = \{(s_t, a_t)\}_{t=0}^{T}$. If the agent has a full observation of the grid, POMDP is referred to as MDP.

Models of Acting Agents in the Gridworld

We consider three different types of acting agents that play in the gridworlds: *Random*, *Reinforcement Learning* (RL), and *Instance-based Learning* (IBL) agents.

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

A **RL** agent uses a tabular form of *Q-learning* algorithm, a quintessential temporal difference approach (Sutton, Barto, et al., 1998). In general, the goal of the RL agent $\mathcal{A}_k$ 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 the 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, $\pi_k^*(s) = \text{argmax}_a Q^*(s,a)$. 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.

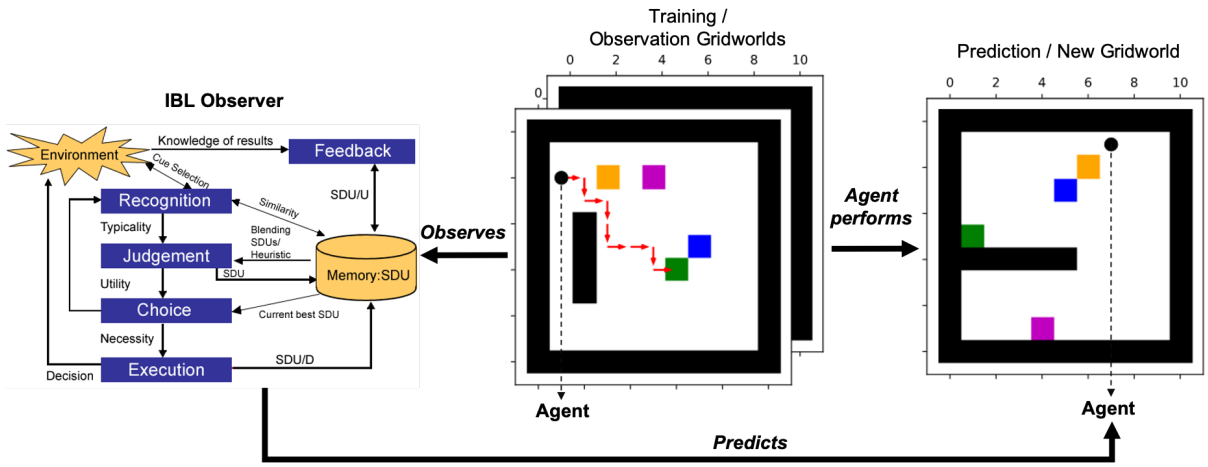


Figure 1: CogToM framework

An **IBL** agent uses the memory and learning mechanisms in IBLT. However, we also explore a currently under-studied situation in IBL models, wherein feedback is sparse and delayed. In the gridworld task, the representation of an *instance* is defined by 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 certain grid. When making a prediction about which action a the agent $\mathcal{A}_k$ will take at state s, the IBL agent selects the action with the highest expected utility using the *blended* value (Equation 3).

Importantly, the agent only gets the real outcome at the end of each episode, typically entailing a sequence of trials. Thus, the IBL agents must learn to update the expected outcome from the consequent reward or penalty, so that different instances are either reinforced or penalized accordingly. To that end, we employ an exploratory mechanism of delayed feedback in the IBL model, where the actual observed outcome is assigned equally to all actions taken in a trajectory. That is, considering the trajectory $\mathcal{T}_k = \{(s_t,a_t)\}_{t=0}^{T}$ if the $\mathcal{A}_k$ 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'.

IBL Observer

IBL observer is a model that is identical to the IBL agent (i.e., the theoretical principles of the IBL model are the same). However, the IBL observer learns from the observations of the agent's decisions in the gridworld, while the IBL agent actually makes the decisions in it.

An instance in the IBL observer is structured in an identical fashion to the IBL agent. The logic behind the IBL observer model, however, is that it learns from past observations of action traces taken by agents in the gridworld, in order to infer and predict the agent's behavior in the *new* gridworld.

The "past experience" of the IBL observer is implemented as proposed in (Lejarraga, Dutt, & Gonzalez, 2012; Gonzalez & Dutt, 2011): inserting "pre-populated instances" in the model's memory. The pre-populated instances correspond to the sequence of decisions the agents made in multiple episodes. More precisely, each observed trajectory $\mathcal{T}_{kj}$ produced by an agent $\mathcal{A}_k$ following its policy π_k in POMDP $\mathcal{M}_j$ is structured as pre-populated instances in the IBL observer's memory. Presumably, each agent has their true reward signal R_k that defines their goal and desire (and that is 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 derived from the inference. This differentiates our work from the approach of Inverse Reinforcement Learning (Ng, Russell, et al., 2000) which is merely aimed at finding a reward function that explains the given agent's history of behavior.

Experiments

To evaluate whether the IBL observer model is able to develop ToM (i.e., the ability to predict desires and beliefs of agents in new gridworld environments), we conducted three experiments including: (1) an *arbitrary goal*, (2) a *goal-directed* task, and (3) a robust test of ToM: *false beliefs*.

It is important to emphasize that in these experiments none of the parameters of any of the models were optimized in any way. The parameters of the agents' models and those of the observer IBL model were "default" values, commonly used in the literature. The IBL observer's parameters were $\sigma = 0.25$ and $d = 0.5$ values that come from the ACT-R cognitive architecture (Anderson & Lebiere, 2014).

Experiment 1: Arbitrary Goal with Random Agents

In this experiment, random agents aim at obtaining one of the four colored objects within a 31 step limit. We created different types of random agents based on their action strategy, i.e. their policy π_k. In turn, given behavioral trajectories of

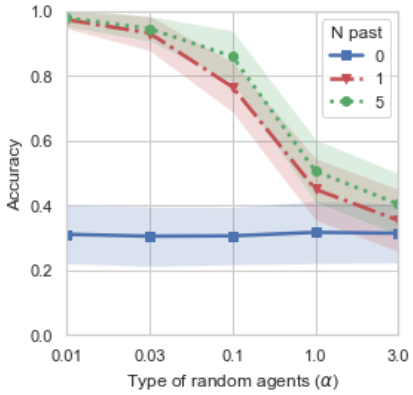


Figure 2: Accuracy of the IBL observer's prediction about the Random agents' initial actions

these random agents on randomly generated gridworlds, the IBL observer was tasked with predicting the initial action that each of the random agents in a new gridworld.

More concretely, we defined different random agents by varying concentration parameter α in each agent's policy that was drawn from a Dirichlet distribution $\pi_k \sim Dir(\alpha)$. If α is close to 0 then the policy of an agent is characterized to be near deterministic. Take, for instance, the agent with $\pi_k \sim Dir(\alpha = 0.01)$, it belongs to the class of agents that is very likely to head in one specific direction (either up, down, left or right). Conversely, if $\pi_k \sim Dir(\alpha = 3)$ then the characteristic of the agent's type is far more stochastic.

We trained an IBL observer by letting it observe the trajectory of the corresponding agents that were randomly generated in various POMDPs. We manipulated the number of past gridworlds (N_{past}) from which the observer could learn to evaluate its performance.

Experimental Setup. We considered five alternative values of $\alpha = \{0.01, 0.03, 0.1, 1, 3\}$ and $N_{past} = \{0, 1, 5\}$. The number of observed agents for each type is 100, and the observer was trained for each type of agent separately on N_{past} gridworlds. Then, given an agent's position in a new gridworld, the IBL observer was queried about that agent's next action. There was no reward function for the random agents as consuming any of the four objects terminated the episode. The accuracy was measured by the proportion of the accurately predicted actions relative to the agent's true next action.

Results. The average of 100 random agents of each type are reported in Fig 2. When $N_{past} = 0$, the curve is flat and nearly constant over the different types of agents since the observer's prediction is independent of α. In contrast, the IBL observer's accuracy immediately increases as the number of past observations increases to $N_{past} = 1$ and 5. It is easier for the IBL observer to predict the agents' behavior with near deterministic policies and the accuracy diminishes as the value of α increases.

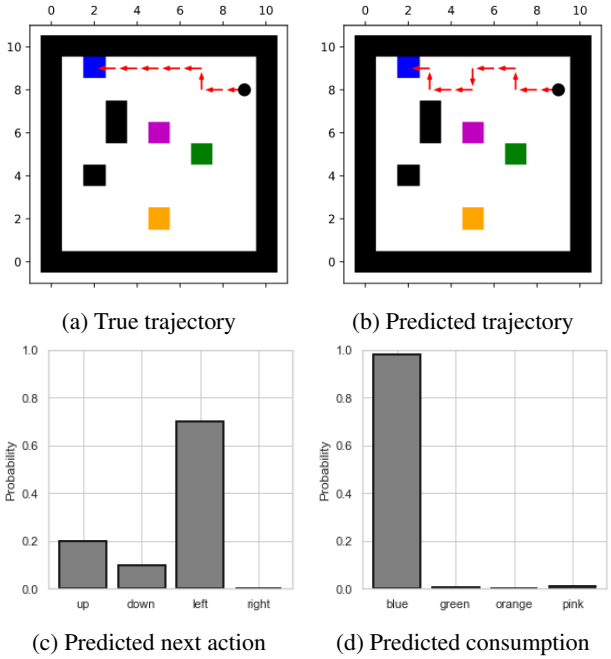
Experiment 2: Goal-Directed Task with RL Agents

The task is set such as that a RL agent is driven by a goal or reaching a particular object that has the highest reward within 31 steps. Consuming any of the other objects results in the termination of the episode.

The IBL observer was required to predict the RL agent's behavior in a new world, given either full or partial observation of the agent's trajectory in a randomly generated training gridworld (MDP). It is important to stress that even though the RL agent's behavior was observable, its reward function along with its policy were completely unknown to the IBL observer. Hence, the IBL observer's mission is to learn to infer which object the agent desires to consume, which is determined by its reward function and transferable to a new gridworld, and then to make behavioural predictions of the agent in the new environment.

We first experimented with the case when the IBL observer was provided with an RL agent's full trajectory in the past gridworld (i.e., *full information*) and then we inspected how the IBL observer performed when it was limited to observing only a partial trajectory (i.e a single action pair in the past MDP, *partial information*).

We analyzed the association between IBL observer's accuracy with full or partial information, varying the number of past MDPs (N_{past}), to assess the IBL observer's predictions about the RL agent's behavior in a new gridworld.



(a) True trajectory (b) Predicted trajectory

(c) Predicted next action (d) Predicted consumption

Figure 3: Illustration of the IBL observer's predictions about the behaviors of goal-directed agents

Experimental Setup. The discount factor γ and learning rate in the RL agent's model were set to 1 and 0.1. Additionally, each agent $\mathcal{A}_k$ was driven by a fixed reward function that rewards the agent with $r_{k,o} \in (0, 1)$ for consuming an ob-

ject o. The vector r_k was drawn from a Dirichlet distribution with concentration parameter $\alpha = 0.01$, which signified that the agent $\mathcal{A}_k$ was favourably attracted to a specific object.

The agent was also penalized for each move and for walking into a wall, 0.01 and 0.05, respectively. Given that the *Q-learning* algorithm (i.e. the RL agent) only converges to optimal values under specific conditions (Dayan & Watkins, 1992), and that to be more in line with human reasoning, the object that the agent consumed by following its sub-optimal policy with respect to the predefined reward function r_k was considered the ground truth. The agent is arguably not always successful in obtaining its true goal (i.e., the object with the highest reward) even though it was trained to be competent at the task after a certain number of episodes, e.g. 500.

As the agent's reward function was concealed from the observer, only the *observable* actions were used for the IBL observer's training. In case the agent's full trajectory was provided, the IBL observer derived from the probability distribution over the objects consumed by the agent in the past MDP to learn about the agent's reward function. When only a single action pair was provided (partial information), the IBL observer was trained to identify the preferred object based on the movement direction of the agent's action.

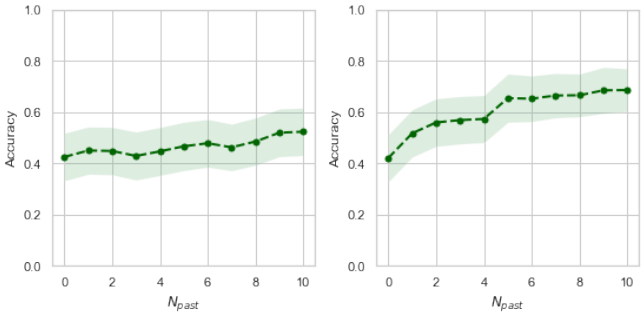
Given the initial location of an agent in a new gridworld, the IBL observer was queried about: (1) the next-step action that the RL agent would take, and (2) the object the RL agent would consume by the end of the episode. We measured the difference between the RL agent's true behaviors (the ground truth) to the IBL observer's predictions. For the analysis of partial trajectories, the value N_{past} was varied from 0 to 10. The experiment was run on different 100 RL agents, and then, we averaged the prediction accuracy over these agents.

Results. Fig 3a and 3b illustrate that the IBL observer's predicted trajectory of the RL agent in a new gridworld is qualitatively aligned with the true trajectory of the agent. Fig 3c and 3d show the IBL observer's predictions of the agent's next action and the object consumption. In the new gridworld the IBL observer predicts the probability of taking the action "left" with about 70% accuracy, and the consuming of "blue" object with about 98% accuracy. The average results from the 100 agents show a mean accuracy in predicting next action is 0.515 ± 0.08 and the goal consumption is 0.687 ± 0.09 with 95% confidence level.

Regarding partial trajectories, Fig 4a and 4b demonstrate that increasing N_{past} can lead to the improvement in the IBL observer's prediction accuracy of the next-step action and of the intended goal.

Experiment 3: False-belief Test with three Agents

Similar to Rabinowitz et al. (2018) we tested the IBL observer for the recognition of agents' *false beliefs* using the *Sally Anne test* (Baron-Cohen, Leslie, & Frith, 1985; Goodman et al., 2006). The Sally-Anne test maps onto the gridworld setting as follows (Table 1): we generated a set of gridworlds in which an agent $\mathcal{A}_k$ was trained to be a blue-object-preferring, but it was required to reach a subgoal (i.e. an additional object) first before returning to consume its preferred blue object. During this time, the location of the preferred object was swapped. Eventually the IBL observer was asked to predict whether or not the agent $\mathcal{A}_k$ would return to the original location of the blue object.



(a) Predicted next action (b) Predicted consumed object

Figure 4: Accuracy of the IBL observer's prediction about the RL agents' behavior when varying N_{past}

Table 1: An overview of simulation design

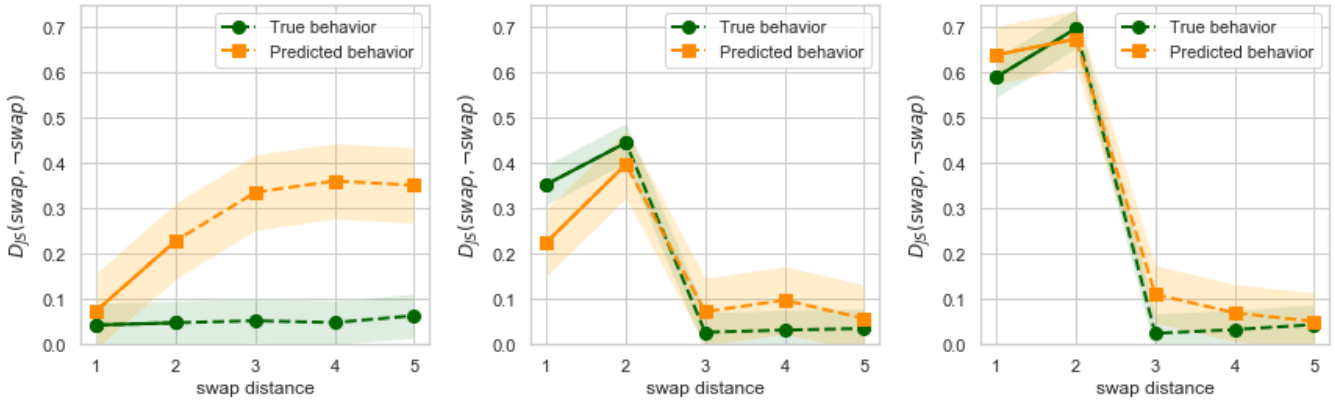
Sally-Anne test	Gridworld task
a) Sally places a marble in a basket	a) An agent $\mathcal{A}_k$ is trained to be a blue-object-prefereing agent
b) Sally moves away	b) $\mathcal{A}_k$ is forced to reach a subgoal
c) Anne puts the marble to a box	c) The location of the preferred object is swapped
d) Where will Sally look for her marble when returning (the basket or the box)?	d) At the subgoal, where will $\mathcal{A}_k$ go to find the preferred blue object (its original or new location)?

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 $\mathcal{A}_k$ sees the swap then $\mathcal{A}_k$ 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.

Importantly, in this test we considered three kinds of agents: Random, RL, and IBL agents. We included an IBL model as an agent to explore whether the IBL observer would be more accurate in developing ToM of an IBL agent than of other models (RL or Random agents) that, by definition, are less aligned with the IBL observer's beliefs.

Experimental Setup. We examined the effect of the swap event on the behavior of the three types of agents, and on the IBL observers' performance. We compared how the 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.

When the swap event occurred, the locations of the four objects were randomly permuted. Moreover, we introduced a distance variable (*dist*) to control whether or not an agent sees the change. If the swap occurred within the agent's view



(a) Random agents (b) RL agents (c) IBL agents

Figure 5: Effect of swap events on the agents' true behavior and on the prediction of IBL observer about the agents' behavior

(i.e. the distance between the agent's and the preferred object's location is within a 2-block radius), the agent' policy was updated according to the change. Conversely, if the swap was outside the agent's view, its policy remained unchanged, which exhibits a sign of a false belief. The agent was rewarded with 1 for consuming the subgoal and a particular preferred object (e.g. the blue object in this experiment).

Since an agent was tasked with consuming the subgoal first and then the preferred object, only the agent's policy in the episodes in which such 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 episodes (e.g. 500). For instance, if the swap happened within the agent's sight then it was less likely to return to the original location (the low frequency). In contrast, if the swap event occurred out of the agent's view then the frequency of revisiting the original position was high.

To evaluate the impact of the swap event on the agent's policy, we used the Jensen-Shannon divergence (D_{JS}) between the probability distribution over the locations associated with the four objects that the agent consumed at the end of the episode in the swap and no swap conditions. Basically, 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 swaps would affect the IBL observers' prediction about the agents' behavior.

Results. Fig 5 shows the performance of each of the three types of agents: Random, RL and IBL agents (100 different agents of each type). As we observed, IBL agents outperform the RL and Random agents in distinguishing the absence and the presence of the swap event when it is visible to the agent ($dis \leq 2$) (solid line section between swap distance 1 and 2). In particular, when the swap event occurs within the agents' view, the IBL 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 different. When $dist > 2$, by contrast, the swap event is invisible to the agent (dot lines), and hence the agent is unable 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.

Fig 5 further reports the results obtained from the IBL observer, when observing 100 agents for each type of Random, RL, IBL agents in terms of D_{JS}. The IBL observers can make the predictions that qualitatively resemble the RL and IBL agents' true behaviors. This, however, does not hold for the Random agents since when the swap occurred but not visible to the agents, the Random agents still were less likely to turn back to the original location due to their random characteristics. As a result, the IBL observer mistakenly learned that the agents saw the swap so they moved away, which results in the increase of D_{JS}.

Finally, we measured the differences between the predictions of the IBL observer about the agents' behavior in terms of Root Mean Square Error (RMSE). The RMSE in predicting the Random, RL and IBL agents' actions is 0.242, 0.071 and 0.048, respectively. The result corroborates our hypothesis that the IBL observer can provide better predictions about the IBL agents than other agents.

Conclusions

We introduce CogToM, that uses a cognitive model generated from IBLT (Gonzalez et al., 2003) to demonstrate the development of ToM from observation of actions of acting agents. This is an advancement over current models of ToM, given that IBL models are cognitively plausible and rely of a

generic theory of decisions from experience. Standard computational models of ToM often make unrealistic assumptions about the rationality of the agents (Baker et al., 2017) and require of complex architectures or complex Machine Learning approaches (Rabinowitz et al., 2018).

We demonstrate a memory-based inference process that uses simple cognitive mechanisms derived from theoretical principles of human cognition. The advantages of using a theoretically-grounded approach are that we are able to explain human inductive learning processes without the need of relying on unrealistic assumptions of human rationality, large amounts of data, or complex models.

Results from our experiments illustrate the ability of the IBL observer to predict next action, beliefs and false beliefs in novel situations after minimal observations of the actions of other agents. Interestingly, an IBL observer is able to predict false beliefs of an IBL agent better than the false beliefs of random and RL agents. Given the recognized ability of IBL models to replicate human behavior (Gonzalez & Dutt, 2011; Gonzalez, 2013; Hertwig, 2015; Lejarraga et al., 2012), this result suggests the IBL model would be able to predict the acting agents' beliefs and actions in similar ways as humans would, although this demonstration is left for future research.

Acknowledgments

This research is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), award number: FP00002636. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

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

Baker, Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Banks, J. (2019). Theory of mind in social robots: Replication of five established human tests. *International Journal of Social Robotics*, 1–12.

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

Botvinick, M., Barrett, D. G., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., ... others (2017). Building machines that learn and think for themselves: Commentary on lake et al., behavioral and brain sciences, 2017. *arXiv preprint arXiv:1711.08378*.

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 *Progress in brain research* (Vol. 202, pp. 73–98). Elsevier.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review*, *118*(4), 523–51.

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, Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... 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).

Hertwig, R. (2015). Decisions from experience. *The Wiley Blackwell handbook of judgment and decision making*, *1*, 240–267.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

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.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml* (Vol. 1, p. 2).

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., & Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning* (Vol. 2) (No. 4). MIT press Cambridge.

Turing, A. (1950). Mind. *Mind*, *59*(236), 433–460.

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.