Cognitive Machine Theory of Mind

Thuy Ngoc Nguyen (ngocnt@cmu.edu) & Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory - Social and Decision Sciences Department
Carnegie Mellon University, Pittsburgh PA 15213 USA

Introduction

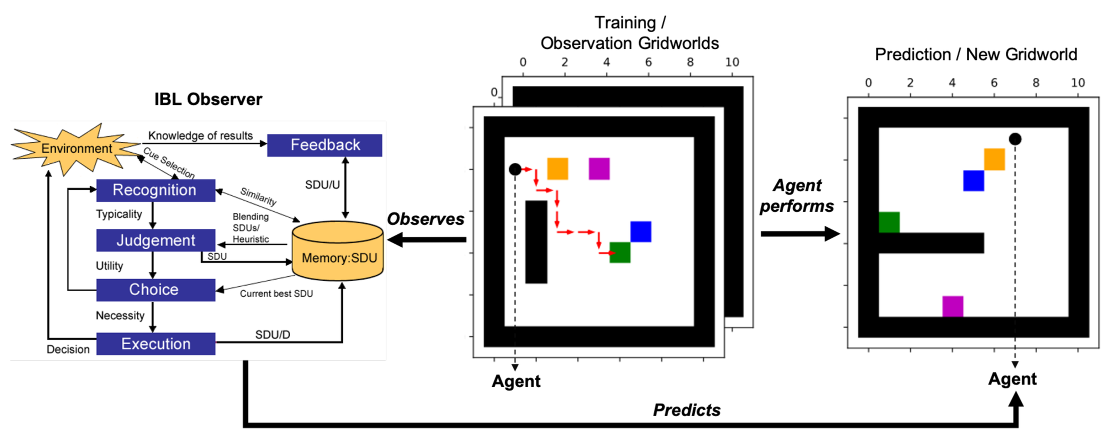
- Theory of mind (ToM): humans' ability to infer and understand the beliefs, desires, and intentions of others [4].
- Cognitive Machine Theory of Mind (CogToM): a framework that relies on a general cognitive theory of decisions from experience, Instance-Based Learning Theory (IBLT) [3].

Instance-Based Learning Theory

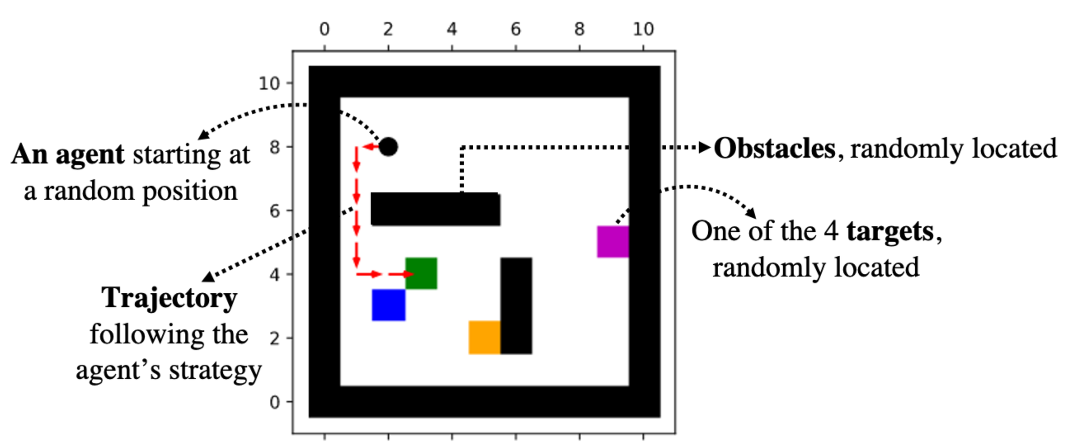
- IBLT explains human learning in dynamic decision problems [3].
- An "*instance*", a memory unit in IBLT, is represented by three elements: a situation (S) (or state s); a decision (D) (or action a taken in state s); and a utility (U) (expected utility or experienced outcome of the taken action taken in a state).
- IBLT uses the Activation equation of the ACT-R architecture [1] for representing how readily available the information is in memory.

CogToM: A Cognitive Machine Theory of Mind Framework

- An *observer* is a cognitive model based on IBLT [3] that builds ToM by observing the actions of *agents* playing in a gridworld.
- The IBL observer can predict the agent's future behavior, such as a next-step action or the agent's desired target in a new gridworld.



- A gridworld is a sequential decision making problem wherein an agent moves through a $N \times N$ grid ($N = 11$) by making decisions (i.e., up, down, left, right) to search for targets.



Models of Acting Agents in the Gridworld

- **Random** agent $\mathcal{A}_k$ selects an action a in state s based on the probability $\pi_k(a|s)$ drawn from a Dirichlet distribution $\pi_k \sim Dir(\alpha)$.
- **Reinforcement Learning (RL)** agent adopts a tabular form of *Q-learning* algorithm, a well-known temporal difference approach [6].
- **Instance-based Learning (IBL)** agent uses the memory and learning mechanisms in IBLT. It selects the action with the highest expected utility using the blended value.

IBL Observer

- Derived from the observable actions of the agent, the IBL observer infers the agent's true reward function.
- Based on the inferred reward, the IBL observer makes the prediction about the agent's behavior in the new environment.
- The "past experience" of the IBL observer is implemented by inserting "pre-populated instances" in the model's memory [2].

Experiments

Following [5], three experiments were conducted: (1) an *arbitrary goal* task, (2) a *goal-directed* task, and (3) a *false beliefs* test of ToM.

Experiment 1: Arbitrary Goal with Random Agents

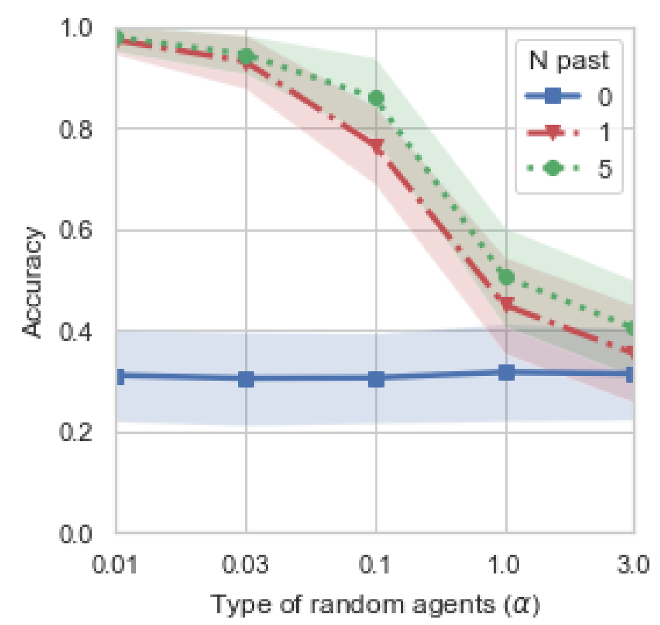
- Agents' goal: obtain one of the four colored objects within 31 steps.
- IBL observer's goal: predict the initial action of the random agents in a new gridworld, given the agents' trajectories in a past gridworld.

Experimental Setup.

- Different types of random agents: $\alpha = \{0.01, 0.03, 0.1, 1, 3\}$
- Different number of past gridworlds: $N_{past} = \{0, 1, 5\}$.
- Number of observed agents for each type is 100.
- Evaluation metric: the proportion of the accurately predicted actions relative to the agent's true next action.

Results.

- $N_{past} = 0$: the observer's prediction is independent of α.
- $N_{past} = 1$ and 5: the IBL observer's accuracy increases.
- Accuracy diminishes as α increases: it is easier for the IBL observer to predict the agents' behavior with near deterministic policies.



Experiment 2: Goal-Directed Task with RL Agents

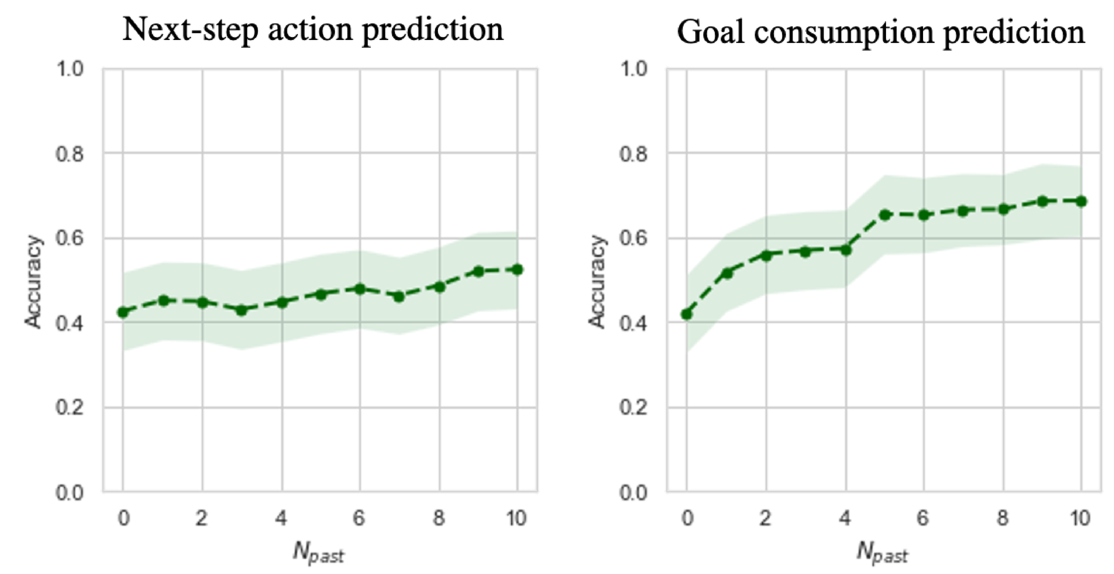
- Agents' goal: obtain a particular object that has the highest reward within 31 steps. Consuming any of the other objects leads to the termination of the episode.
- IBL observer's goal: learn to infer which object the RL agent desires to consume, and then predict (1) the next-step action that the agent would take, and (2) the object the agent would consume in the new environment, given either *full* or *partial* observation of the agent's trajectory in a training gridworld.

Experimental Setup.

- Each agent $\mathcal{A}_k$ is driven by a fixed reward, $r_{k,j} \in (0, 1)$, for consuming an object o_j where $j = 1, \ldots, 4$.
- For the analysis of partial trajectories, $N_{past} = 1..10$.
- Number of RL agents is 100.
- Evaluation metric: the difference between the RL agent's true behaviors (the ground truth) and the IBL observer's predictions.

Results.

- Prediction accuracy: (1) next-step action is 0.515 ± 0.08; and (2) goal consumption is 0.687±0.09 with 95% confidence level.
- Regarding partial trajectories, the IBL observer's prediction accuracy is improved when increasing N_{past}.



Experiment 3: False-belief Test with three Agents

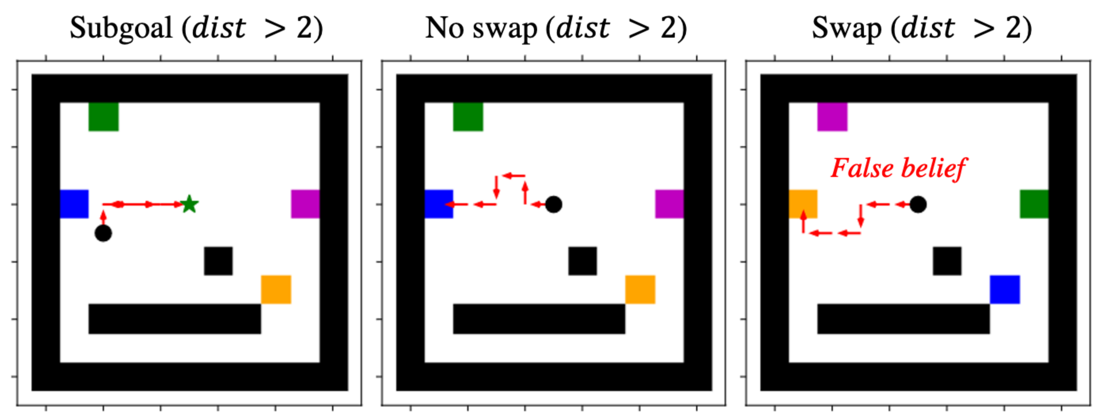
- *Sally Anne* test is mapped onto the gridworld setting as follows:

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)?

- IBL observer's goal: recognize agents' *false beliefs* given the awareness of the changes (i.e. swap event). The IBL observer can indicate:
  - if the agent $\mathcal{A}_k$ sees the swap then $\mathcal{A}_k$ it will not go back to the original location (a sign of *true belief*).
  - if the agent $\mathcal{A}_k$ is not aware of the swap then it will return to the original location (a sign of a *false belief*).

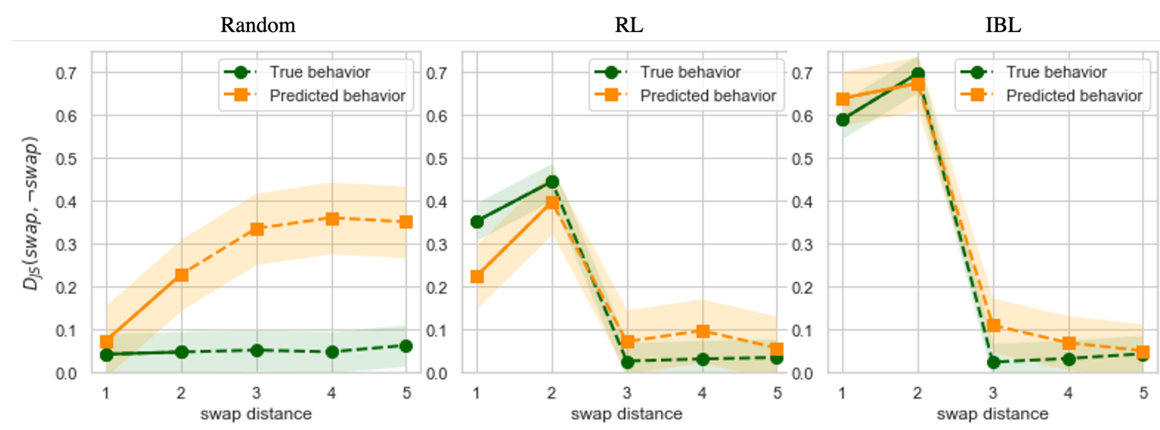
Experimental Setup.

- 3 kinds of agents: Random, RL, and IBL (100 agents of each type).
- Field of view: within 2 blocks, i.e. $dist \leq 2$: the agent sees the swap (its policy is updated); $dist > 2$: the agent cannot see the swap.
- Evaluate (1) how the agent behaves in the *swap* and *no swap* settings and (2) how the IBL observer performs when observing different types of agents in the two settings.
- Evaluation metric: Jensen-Shannon divergence (D_{JS}) between the probability distribution over the locations of the four objects that the agent consumed in the *swap* and *no swap* events.



Results.

- IBL agents outperform the RL and Random agents in distinguishing the *swap* and *no swap* events when the swap distance $dist \leq 2$; when $dist > 2$: $D_{JS}(swap, \neg swap)$ close to 0.
- IBL observer can provide better predictions about the IBL agents than about the other agents: RMSE in predicting the Random, RL and IBL agents' actions is 0.242, 0.071 and 0.048, respectively.



Conclusions

- Use the IBL process of IBLT [3] and the formulations of the ACT-R architecture [1] for memory-based inference to demonstrate how ToM develops from observation of other acting agents' actions.
- Illustrate the ability of the IBL observer to predict next-step action, intention, and false beliefs in novel situations.

References

[1] J. R. Anderson and C. J. Lebiere. *The atomic components of thought*. Psychology Press, 2014.

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

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

[4] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

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

[6] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge, 1998.

Acknowledgements

This research is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), award number: FP00002636.