Theory of Mind and AI - II

Swati Mishra

Human Centered Artificial Intelligence

Graduate Course - CAS 783

Winter 2025

McMaster University
ENGINEERING

What is Theory of Mind

- Ability for tracking other people's mental states is known as theory of mind.
- Theory of mind refers to an interconnected set of notions that are combined to explain, predict, and justify the behavior of others
- The understanding that others have intentions, desires, beliefs, perceptions, and emotions different from one's own and that such intentions, desires, and so forth affect people's actions and behaviors.

Machine Theory of Mind

Machine Theory of Mind

Neil C. Rabinowitz [1] **Frank Perbet** [1] **H. Francis Song** [1] **Chiyuan Zhang** [2] **Matthew Botvinick** [1]

Abstract

Theory of mind (ToM) broadly refers to humans' ability to represent the mental states of others, including their desires, beliefs, and intentions. We design a Theory of Mind neural network – a *ToMnet* – which uses meta-learning to build such models of the agents it encounters. The ToMnet learns a strong prior model for agents' future behaviour, and, using only a small number of behavioural observations, can bootstrap to richer predictions about agents' characteristics and mental states. We apply the ToMnet to agents behaving in simple gridworld environments, showing that it learns to model random, algorithmic, and deep RL agents from varied populations, and that it passes classic ToM tasks such as the "Sally-Anne" test of recognising that others can hold false beliefs about the world.

of others, such as their desires and beliefs. This ability is typically described as our Theory of Mind (Premack & Woodruff, 1978). While we may also leverage our own minds to simulate others' (e.g. Gordon, 1986; Gallese & Goldman, 1998), our ultimate human understanding of other agents is not measured by a correspondence between our models and the mechanistic ground truth, but instead by how much they enable prediction and planning.

In this paper, we take inspiration from human Theory of Mind, and seek to build a system which learns to model other agents. We describe this as a *Machine Theory of Mind*. Our goal is not to *assert* a generative model of agents' behaviour and an algorithm to invert it. Rather, we focus on the problem of how an observer could learn *autonomously* how to model other agents using limited data (Botvinick et al., 2017). This distinguishes our work from previous literature, which has relied on hand-crafted models of agents as noisy-rational planners – e.g. using inverse

RL Basics

In Reinforcement Learning, we have an agent that interacts with its environment. At each time step, the agent will get some input from the environment. Then, the agent selects an action to take. The environment is then transitioned into a new state and the agent is given a reward for its action.

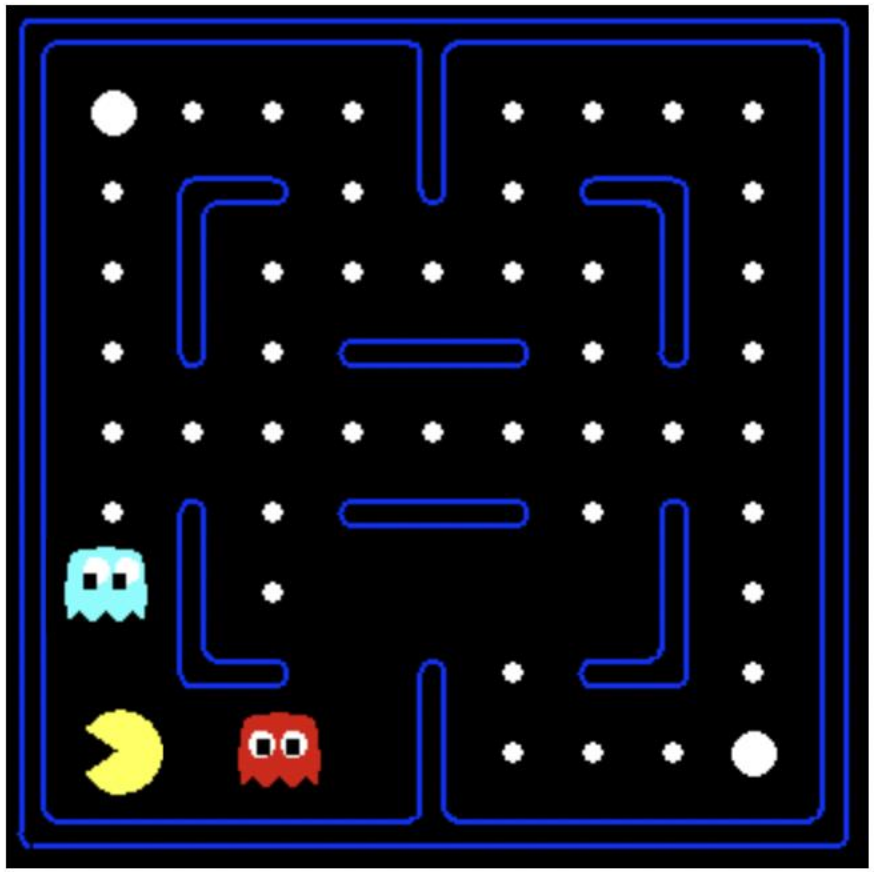
Any RL task can be defined with the following components:

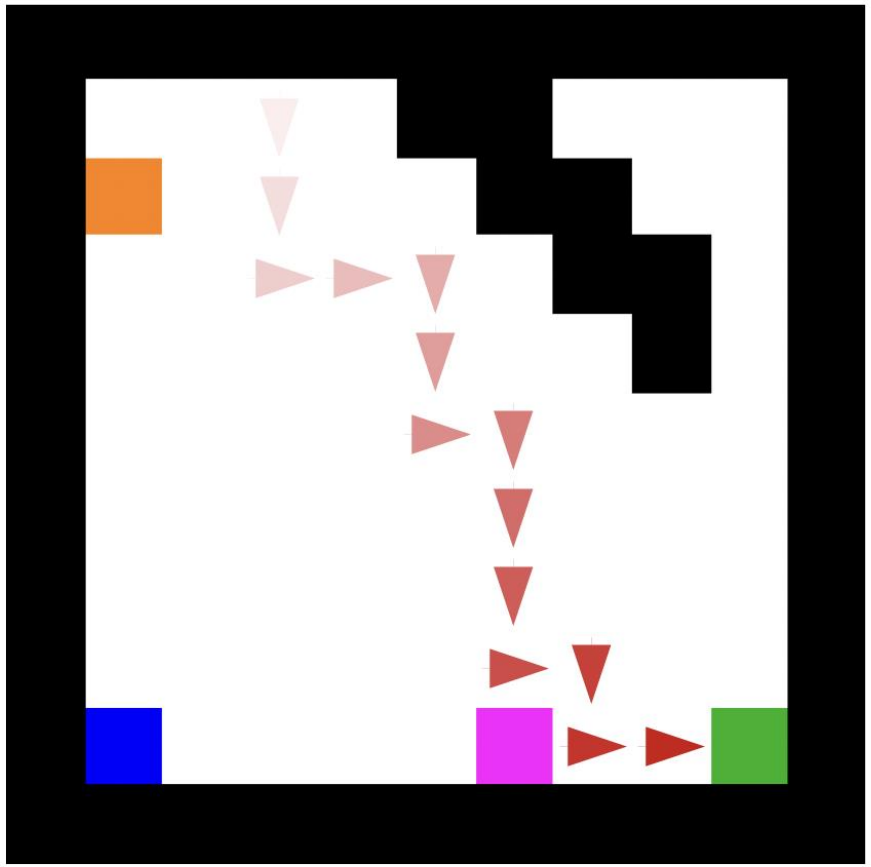
- Agent
- States
- Actions
- Rewards

RL Basics

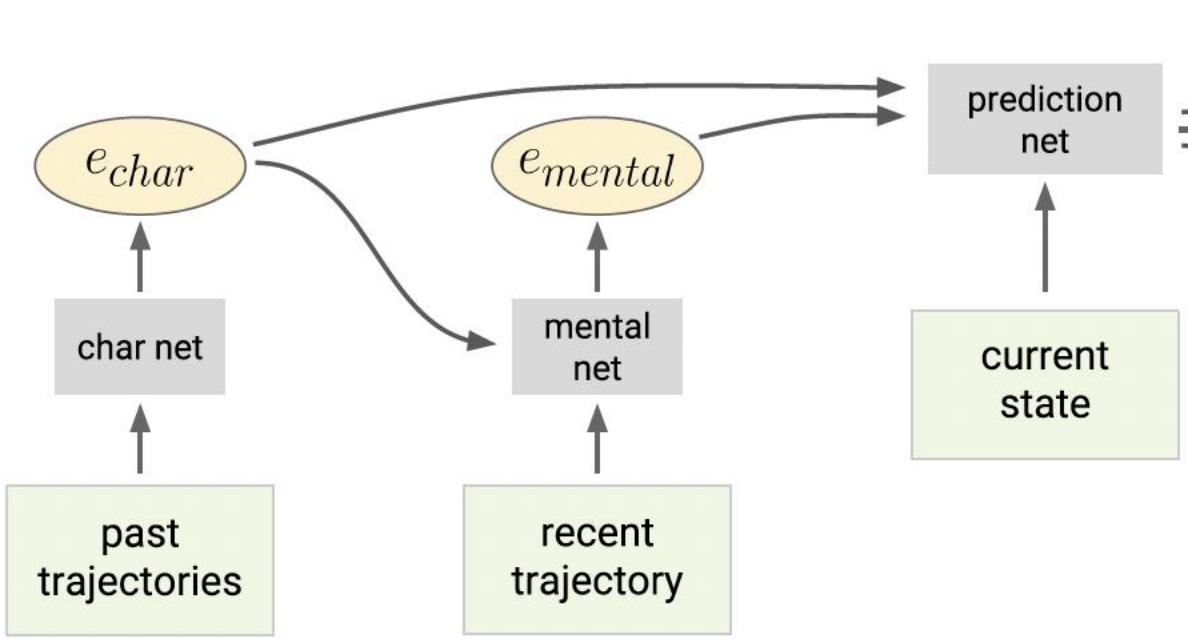
- Markov's property states that the future depends only on the present, not on the past. A Markov chain is a probabilistic model that represent this kind of approach. Moving from one state to another is called transition and its probability is called transition probability
- When a stochastic process follows Markov's property, it is called a Markov Process. MDP is an extension of the Markov chain.
- A MDP has 4 elements:
  - A set of states(S) the agent can be in.
  - A set of actions (A) that can be performed by an agent to move from one state to another.
  - A set of transition probabilities (${P^a}_{ss}{}'$), which define the probability of moving from state s to state s' by performing action a.
  - A set of reward probabilities (${R^a}_{ss'}$), which defines the probability of a reward acquired for moving from state s to state s' by performing action a.

Grid world Task





ToMnet - Architecture



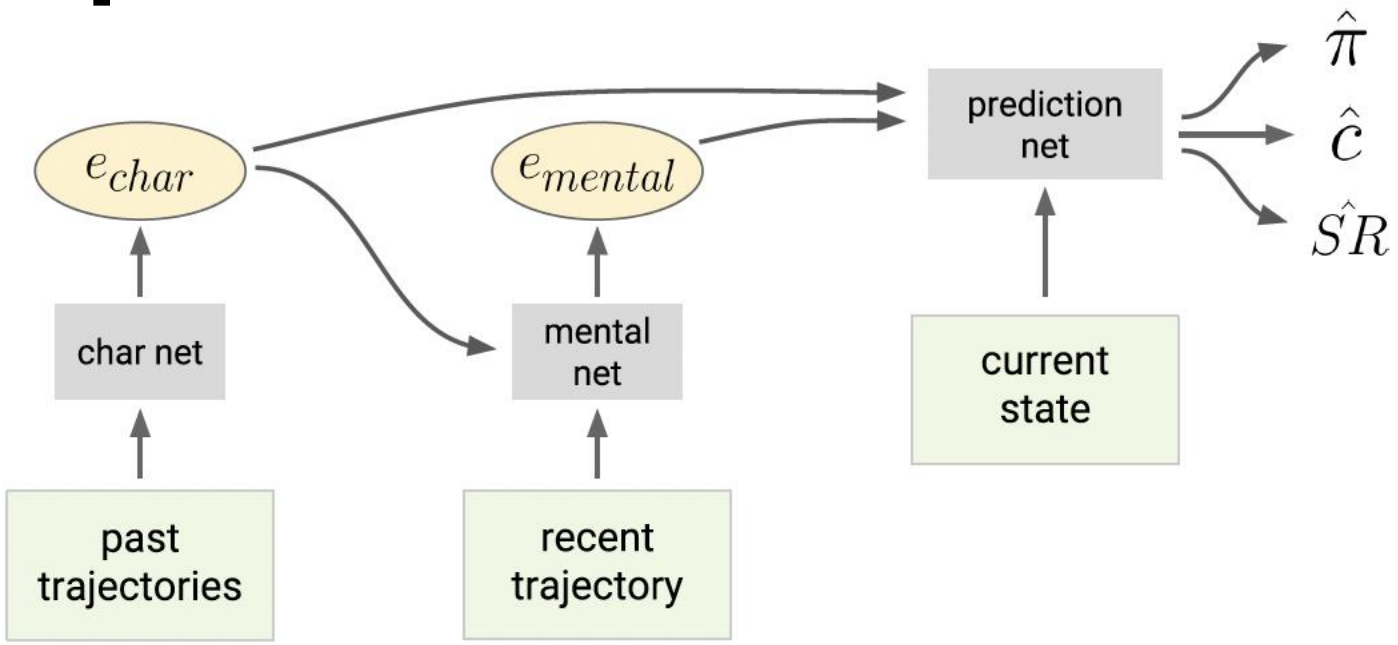
$\hat{\pi}$ next-step actions

$\hat{c}$ whether objects will be consumed

$\hat{SR}$ successor representations

Predictions about future behaviour

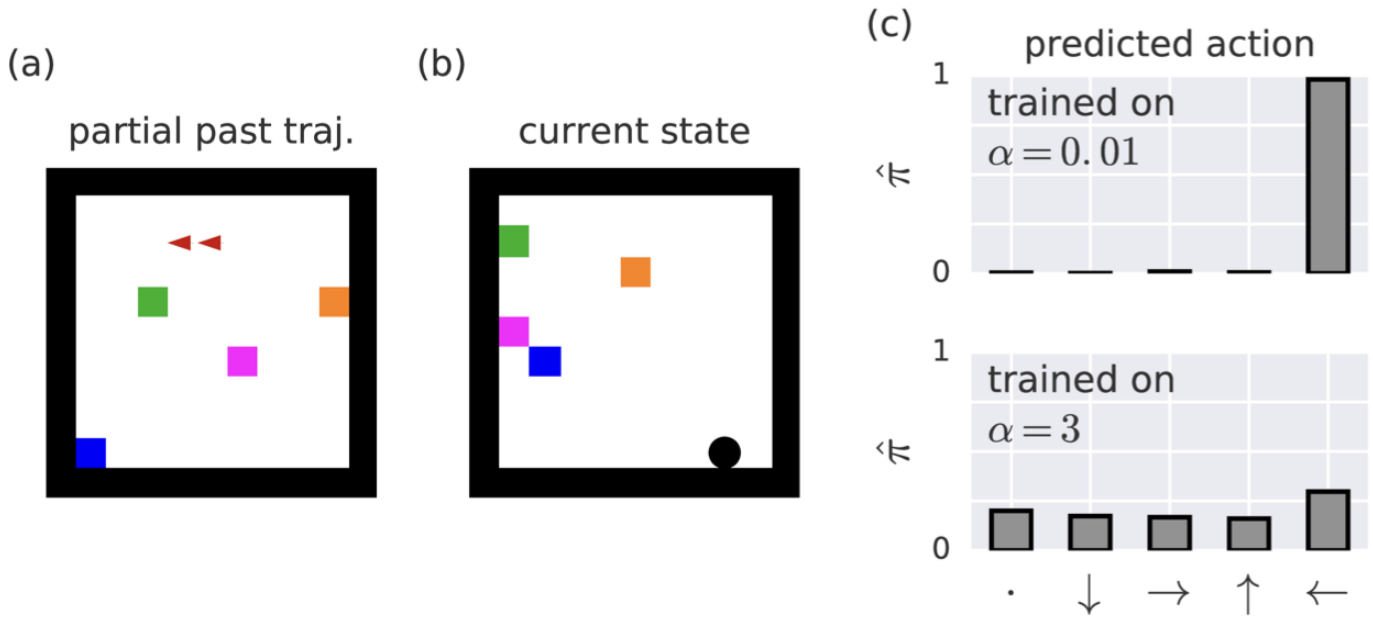
Experiments



- General theory of mind: The learned weights of the network, which encapsulate predictions about the common behavior of all agents in the training set
- Agent-specific theory of mind: The "agent embedding" formed from observations about a single agent at test time, which encapsulates what makes this agent's character and mental state distinct from others'. These correspond to a prior and posterior over agent behaviour.

Experiments

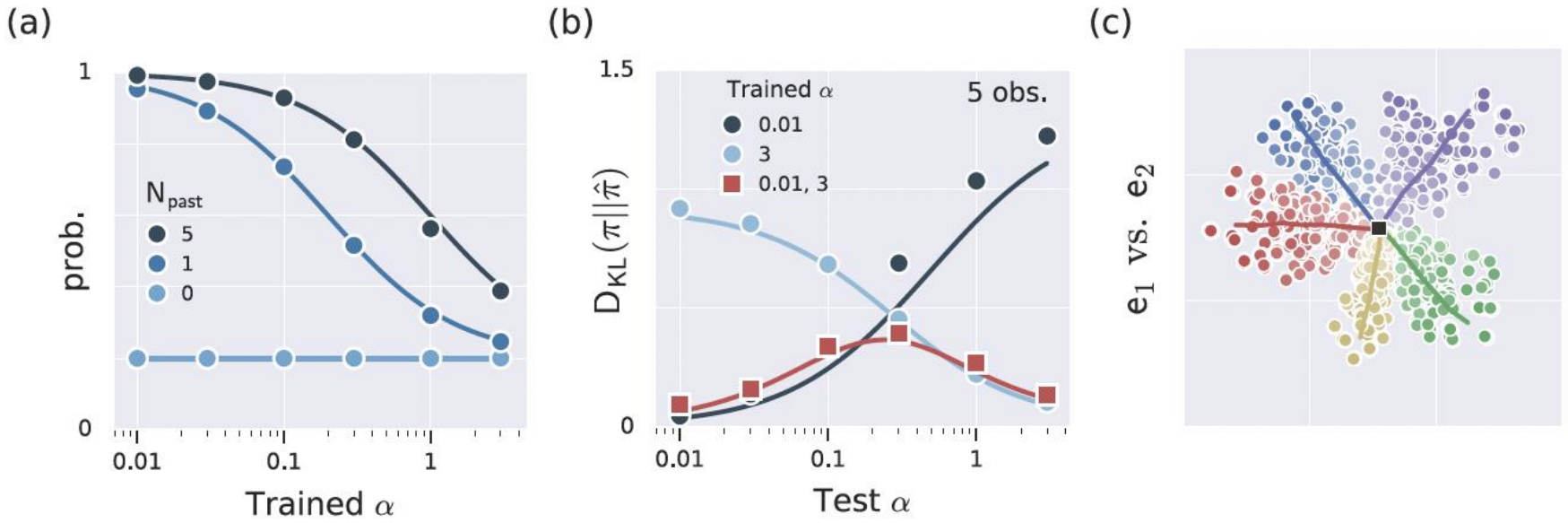
Experiment 1: Random Agents



Increasing the value of alpha makes the policy more stochastic.

Experiments

Experiment 1: Random Agents

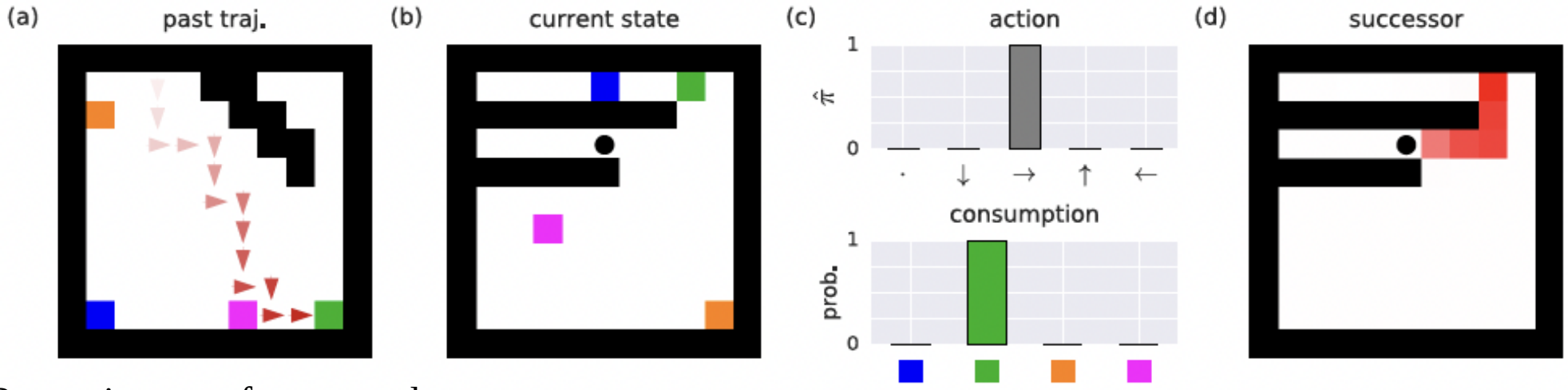


Without any changes to its architecture, a ToMnet learns a general theory of mind that is specialized for the distribution of everyone it encounters, and estimates an agent-specific theory of mind online for each individual.

Experiments

Experiment 2: Inferring goal-directed behavior

Black dot: agent position.



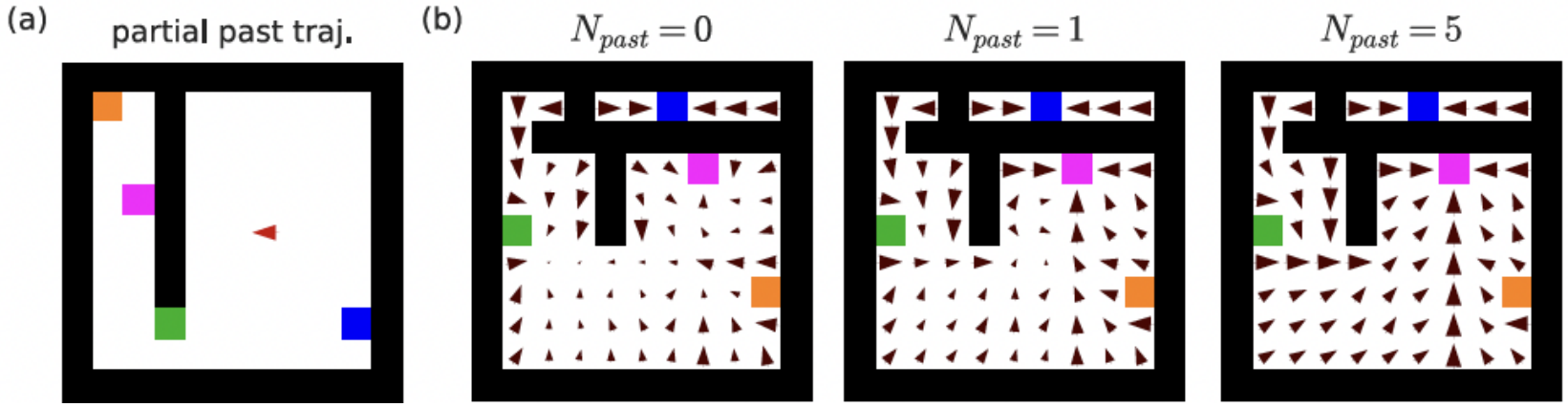
Past trajectory of an example agent. Colored squares: the four objects. Red arrows: agent's position and action.

Example query: a state from a new MDP.

ToMnet's predictions for the query in (b), given the past observation in (a).
SR in (d) for discount = 0:9. Darker shading: higher SR.

Experiments

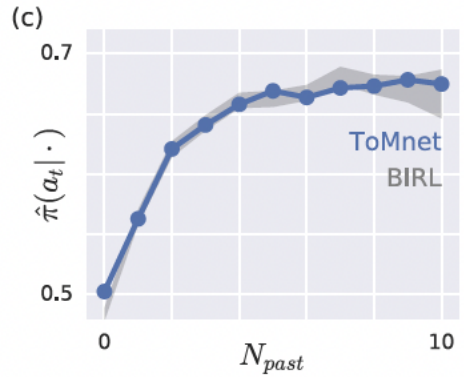
Experiment 2: Inferring goal-directed behavior – where ToMnet sees only a partial snapshot



This ToMnet sees only snapshots of single observation/action pairs (red arrow) from a variable number of past episodes.

Predicted policy for different initial agent locations in aquery MDP. Arrows: resultant vectors for the predicted policies

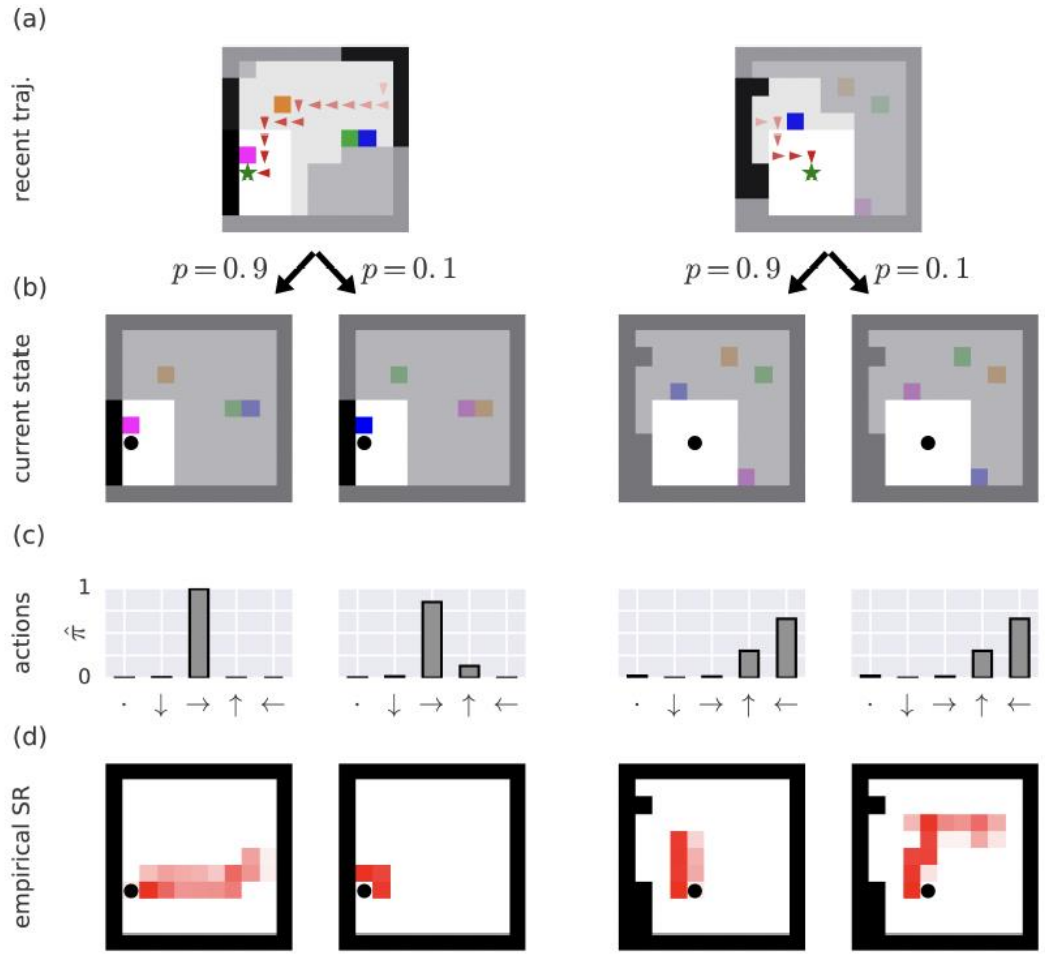
The average posterior probability assigned to the true action



Experiments

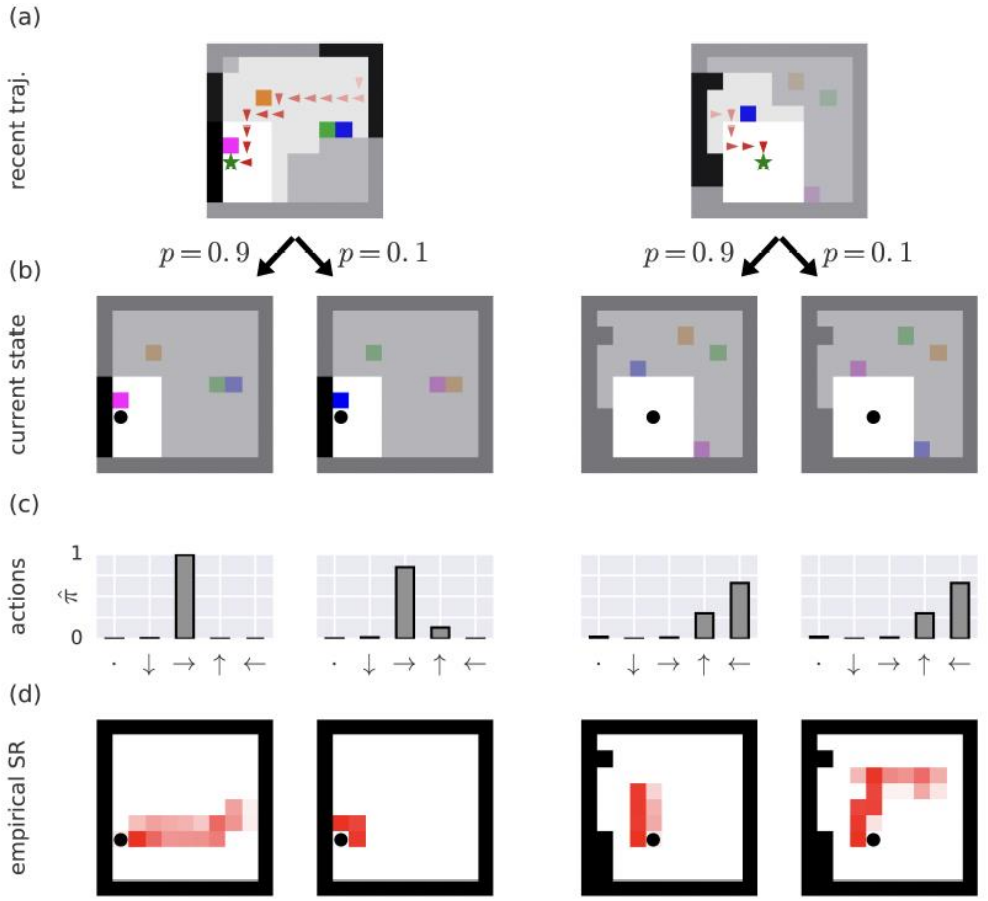
Experiment 3: Subgoal task, where agents can have false beliefs.

The authors introduced random state changes that agents might not see. In the subgoal maze described above, we included a low probability (p = 0:1) state transition when the agent stepped on the subgoal, such that the four other objects would randomly permute their locations instantaneously.



Experiments

Experiment 3: Subgoal task, where agents can have false beliefs.



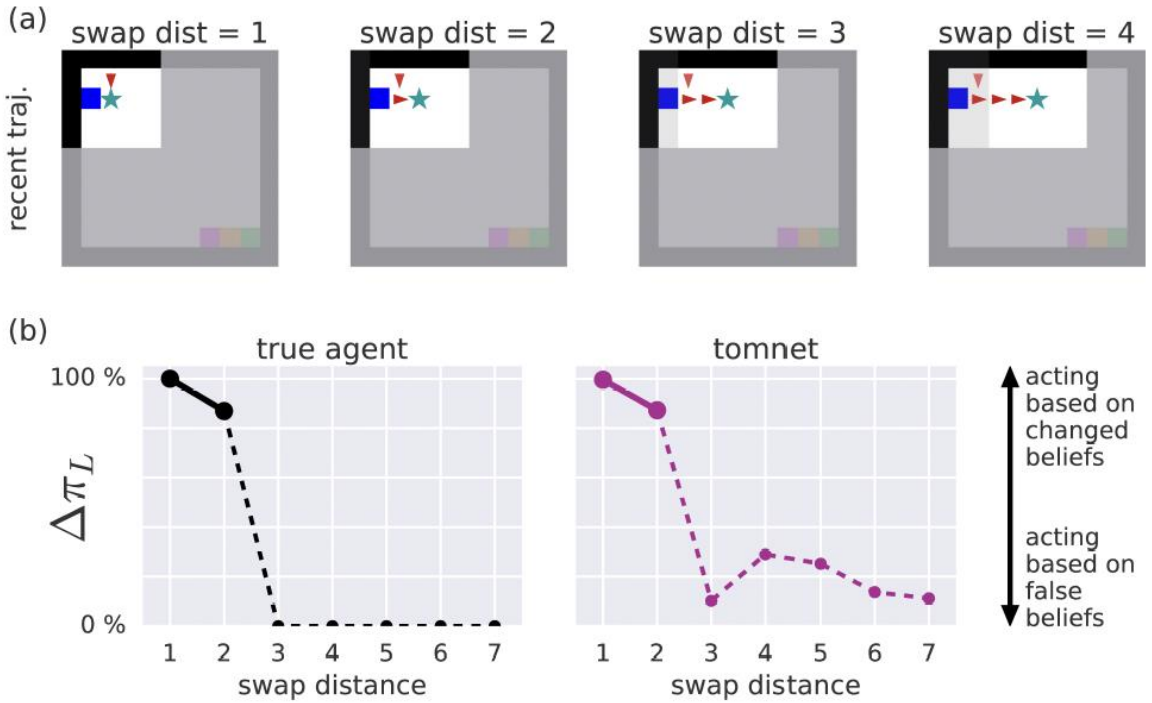
Trajectory of an agent (red arrows) as it seeks the subgoal (star). Agent has partial observability: dark grey areas have not been observed; light grey areas have been seen previously, but are not observable at the time of subgoal consumption. **(b)** When the agent consumes the subgoal object, there is a small probability that the other objects will instantaneously swap locations. Left: swap event within the agent's current field of view. Right: outside it. **(c)** Effect of swap on agent's immediate policy. **(d)** Effect of swap on agent's empirical SR (computed over 200 stochastic rollouts). Agent prefers the blue object.

Experiments

Experiment 3: Acting based on false beliefs: Sally-Anne test.

(a) Agents' forced trajectory. When it reaches the subgoal (star), a swap event may or may not occur. If there is no swap, the optimal action is to go left. By extending the length of the path, the swap event will no longer be visible to the agent.

(b) Left: effect of a swap event on the agents' true policies, measured as the relative reduction in their probability of moving back towards the original location where they saw the blue object. If the agent can see that the object has moved from this location it will not return left. If it cannot see this location, its policy will not change.



Discussion

Experiment

Humanness lies in unpredictability: Role of Theory of Mind on anthropomorphism in human-computer interactions

Julia Ayache
julia.ayache@ntu.ac.uk
Nottingham Trent University
Nottingham, United Kingdom

Andrew M Connor
Auckland University of Technology
Auckland, New-Zealand

Stefan Marks
Auckland University of Technology
Auckland, New-Zealand

Alexander Sumich
Nottingham Trent University
Nottingham, United Kingdom
Auckland University of Technology
Auckland, New-Zealand

Nadja Heym
Nottingham Trent University
Nottingham, United Kingdom

Why do we need to predict

- The capacity to predict the behavior of a virtual agent is a core aspect of human-computer interaction.
- Predictability is desirable and prevents out-of-the-loop problems by allowing humans to monitor and predict the activity of autonomous agents.
- Predictability can be detrimental by decreasing the perceived agency of autonomous agents displaying stereotypical behaviors.
- Predicting a virtual agent's behavior relies on a mechanism similar to that used in predicting human behaviors, such as Theory of Mind abilities.
- The need for predictability is one of the main factors leading to attributing human-like behavior to virtual agents – anthropomorphism (unpredictability is a marker of humanness)

Measurements

- It is expected that higher perspective-taking scores will be associated with better social functioning.
- Perspective-taking capability for nonegocentric behavior— that is, behavior that subordinates the self (or the self's perspective) to the larger society made up of other people.
- Perspective-taking ability should allow an individual to anticipate the behavior and reactions of others, therefore, facilitating smoother and more rewarding interpersonal relationships.

Measure
Interpersonal functioning
Shyness
Loneliness
Social anxiety (SCI)
Audience anxiety
M^- (EPAQ)
F_{VA}^- (EPAQ)
F_C^- (EPAQ)
Extraversion (SM Scale)
Self-esteem
TSBI
Briggs et al. Self-Esteem Scale
Emotionality
M–F scale (PAQ)
Fearfulness (EASI)
Sensitivity to others
Public self-consciousness (SCI)
Other-directedness (SM Scale)
F scale (PAQ)
Intelligence
SAT-Quantitative
SAT-Verbal
WAIS Vocabulary

Subscale
Perspective-Taking
Fantasy
Empathic Concern
Personal Distress

Hypothesis

- H1: A positive association between participants' self-reported Perspective Taking capabilities and effective ToM abilities in predicting virtual agent's behavior during MPG.

- H2: A negative association between virtual agents' predictability and tendencies for anthropomorphism.

Experiment

Matching Pennies Game:

Participants played the Matching Pennies Game (MPG) by guessing the virtual agent's choice (i.e., picking a blue or red card). The bias of the virtual agent was manipulated across four conditions using a within-subject design, by increasing the tendency to choose a specific card color from 50% to 80% in increments of 10%. The level of bias was randomized within-subject. For each correct answer (i.e., predicting the card accurately that the virtual agent will pick up), the participant won 1 point.

Participants:

Participants (N = 38, 15 men, 21 women and 2 non-binaries, mean age = 22.13, SD= 4.23 years) were recruited from a student population enrolled at Nottingham Trent University.

Results

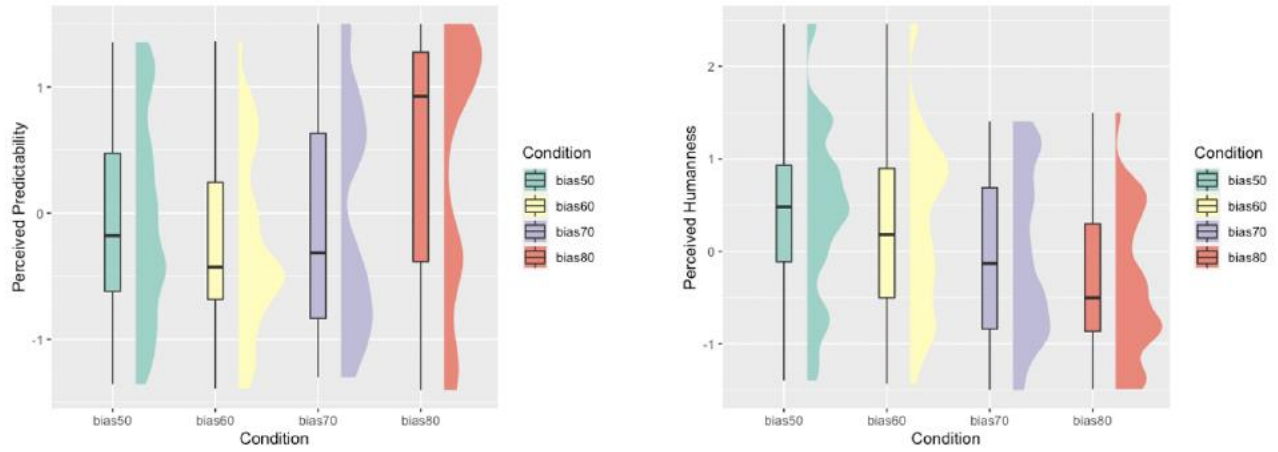


Figure 1: Raincloudplot for perceived predictability (left) and perceived humanness (right) and experimental conditions.

Discussion

Thank you