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.

1. Introduction

What does it actually mean to "understand" another agent? As humans, we face this challenge every day, as we engage with other humans whose latent characteristics, latent states, and computational processes are almost entirely inaccessible. Yet we can make predictions about strangers' future behaviour, and infer what information they have; we can plan our interactions with others, and establish efficient and effective communication.

A salient feature of these "understandings" of other agents is that they make little to no reference to the agents' true underlying structure. A prominent argument from cognitive psychology is that our social reasoning instead relies on high-level *models* of other agents (Gopnik & Wellman, 1992). These models engage abstractions which do not describe the detailed physical mechanisms underlying observed behaviour; instead, we represent the *mental states* 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 (Ng et al., 2000; Abbeel & Ng, 2004), Bayesian Theory of Mind (Baker et al., 2011; Jara-Ettinger et al., 2016) or game theory (Yoshida et al., 2008; Camerer, 2010). Likewise, other work, such as Foerster et al. (2017) and Everett (2017), expect other agents to conform to known, strong parametric models, while concurrent work by Raileanu et al. (2018) assumes that other agents are functionally identical to oneself. In contrast, we make very weak assumptions about the generative model driving others' behaviour. The approach we pursue here learns the agent models, and how to do inference on them, from scratch, via meta-learning.

Building a rich, flexible, and performant Machine Theory of Mind may well be a grand challenge for AI. We are not trying to solve all of this here. A main message of this paper is that many of the initial challenges of building a ToM can be cast as simple learning problems when they are formulated in the right way.

There are many potential applications for this work. Learning rich models of others will improve decision-making in complex multi-agent tasks, especially where model-based planning and imagination are required (Hassabis et al., 2013; Hula et al., 2015; Oliehoek & Amato, 2016). Our work thus ties in to a rich history of opponent modelling (Brown, 1951; Albrecht & Stone, 2017); within this con-

An extended version is available at arxiv.org/abs/1802.07740.

[1]DeepMind [2]Google Brain. Correspondence to: Neil Rabinowitz <ncr@google.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

text, we show how meta-learning can provide the ability to build flexible and sample-efficient models of others on the fly. Such models will also be important for value alignment (Hadfield-Menell et al., 2016), flexible cooperation (Nowak, 2006; Kleiman-Weiner et al., 2016), and mediating human understanding of artificial agents.

We consider the challenge of building a Theory of Mind as essentially a meta-learning problem (Schmidhuber et al., 1996; Thrun & Pratt, 1998; Hochreiter et al., 2001; Vilalta & Drissi, 2002). At test time, we want to be able to encounter a novel agent whom we have never met before, and already have a strong and rich prior about how they are going to behave. Moreover, as we see this agent act in the world, we wish to be able to form a posterior about their latent characteristics and mental states that will enable us to improve our predictions about their future behaviour.

We formulate a task for an observer, who, in each episode, gets access to a set of behavioural traces of a novel agent, and must make predictions about the agent's future behaviour. Over training, the observer should get better at rapidly forming predictions about new agents from limited data. This "learning to learn" about new agents is what we mean by meta-learning. Through this process, the observer should also acquire an effective prior over the agents' behaviour that implicitly captures the commonalities between agents within the training population.

We introduce two concepts to describe components of this observer network: its *general theory of mind* – the learned weights of the network, which encapsulate predictions about the common behaviour of all agents in the training set – and its *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.

This paper is structured as a sequence of experiments of increasing complexity on this Machine Theory of Mind network, which we call a *ToMnet*. These experiments showcase the idea of the ToMnet, exhibit its capabilities, and demonstrate its capacity to learn rich models of other agents incorporating canonical features of humans' Theory of Mind, such as the recognition of false beliefs.

Several experiments in this paper are inspired by the seminal work on Bayesian Theory of Mind (Baker et al., 2011; 2017). We do not try to directly replicate the experiments in these studies, which seek to explain human judgements in computational terms. Our emphasis is on machine learning, scalability, and autonomy. While our work generalises many of the constructions of those previous experiments, we leave the precise alignment to human judgements as future work.

2. Model

2.1. The tasks

We assume we have a family of partially observable Markov decision processes (POMDPs) $\mathcal{M} = \bigcup_j \mathcal{M}_j$. Unlike the standard formalism, we associate the reward functions, discount factors, and conditional observation functions with the agents rather than with the POMDPs. For example, a POMDP could be a gridworld with a particular arrangement of walls and objects; different agents, when placed in the same POMDP, might receive different rewards for reaching these objects, and be able to see different amounts of their local surroundings. We only consider single-agent POMDPs here; the multi-agent extension is simple. When agents have full observability, we use the terms MDP and POMDP interchangeably.

Separately, we assume we have a family of agents $\mathcal{A} = \bigcup_i \mathcal{A}_i$, with their own observation spaces, observation functions, reward functions, and discount factors. Their policies might be stochastic (Section 3.1), algorithmic (Section 3.2), or learned (Sections 3.3–3.5). We do not assume that the agents' policies are optimal for their respective tasks. The agents may be stateful, though we assume their hidden states do not carry over between episodes.

The POMDPs we consider here are 11×11 gridworlds with a common action space (up/down/left/right/stay), deterministic dynamics, and a set of consumable objects (see Appendix C). We experimented with these POMDPs due to their simplicity and ease of control; our constructions should generalise to richer domains too. We parameterically generate individual $\mathcal{M}_j$ by randomly sampling wall, object, and initial agent locations.

In turn, we consider an observer who makes potentially partial and/or noisy observations of agents' trajectories. Thus, if agent $\mathcal{A}_i$ follows its policy π_i on POMDP $\mathcal{M}_j$ and produces trajectory $\tau_{ij} = \{(s_t, a_t)\}_{t=0}^{T}$, the observer would see $\tau_{ij}^{(obs)} = \{(x_t^{(obs)}, a_t^{(obs)})\}_{t=0}^{T}$. For all our experiments, we assume that the observer has unrestricted access to the MDP state and agents' actions, but not to agents' parameters, reward functions, policies, or identifiers.

ToMnet training involves a series of encounters with individual agents. The ToMnet observer sees a set of full or partial "past episodes", wherein a single, unlabelled agent, $\mathcal{A}_i$, produces trajectories, $\{\tau_{ij}\}_{j=1}^{N_{\text{past}}}$, as it executes its policy within the respective POMDPs, $\mathcal{M}_j$. N_{past} might vary, and may even be zero. The task for the observer is to predict the agent's behaviour (e.g. atomic actions) and potentially its latent states (e.g. beliefs) on a "current episode" as it acts within POMDP $\mathcal{M}_k$. The observer may be seeded with a partial trajectory in $\mathcal{M}_k$ up to time t.

The observer must learn to predict the behaviour of *many*

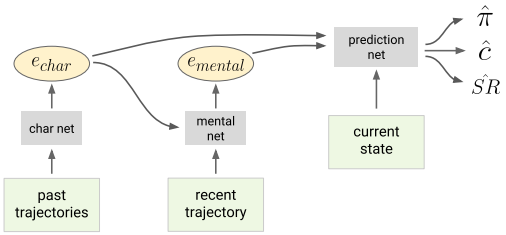

Figure 1. **ToMnet architecture.** Predictions about future behaviour include next-step actions ($\hat{\pi}$), whether objects will be consumed ($\hat{c}$), and successor representations ($\widehat{SR}$; Dayan, 1993).

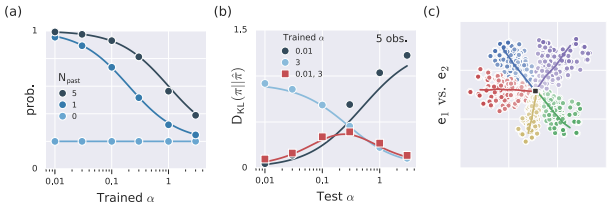

Figure 2. **ToMnet trained on random agents. (a)** Predictions that an agent will perform a particular action, by ToMnets trained on different species, $\mathcal{S}(\alpha)$. Posteriors shown after observing that agent perform just that same action in $N_{\text{past}} = 0$ (prior), 1, or 5 past episodes. Dots: ToMnet predictions. Solid lines: Bayes-optimal posteriors for the respective $\mathcal{S}(\alpha)$. **(b)** Average KL between agents' true and predicted policies when the ToMnet is trained on agents from one species (or mixture), but tested on agents from a different species. Dots and lines as in (a). **(c)** Embeddings $e_{\text{char}} \in \mathbb{R}^2$ of different agents. Colours: most frequent action over 10 past episodes; darker for higher counts.

agents, whose rewards, parameterisations, and policies may vary considerably. This problem resembles one-shot imitation learning (Duan et al., 2017; Wang et al., 2017), but differs since the objective is not for the observer to execute the behaviours itself. Moreover, there is an informational asymmetry, where the "teacher" may conceivably know *less* about the environment state than the "student", and may also carry systematic biases.

2.2. The architecture

To solve these tasks, we designed the *ToMnet* architecture shown in Fig 1. The ToMnet is composed of three modules: a *character net*, a *mental state net*, and a *prediction net*.

The goal of the character net is to *characterise* the presented agent, by parsing observed past episode trajectories, $\{\tau_{ij}^{(obs)}\}_{j=1}^{N_{\text{past}}}$, into a character embedding, $e_{\text{char,i}}$.

The goal of the mental state net is to *mentalise* about the presented agent during the current episode (i.e. infer its mental state; Dennett, 1973; Frith & Frith, 2006), by parsing the current episode trajectory, $\tau_{ik}^{(obs)}$, up to time $t-1$ into a mental state embedding, $e_{\text{mental},i}$. For brevity, we drop the agent subscript, i.

Lastly, the goal of the prediction net is to leverage the character and mental state embeddings to predict subsequent behaviour of the agent. Precise details of the architecture, loss, and hyperparameters for each experiment are given in Appendix A. We train the whole ToMnet end-to-end.

We deploy the ToMnet to model agents belonging to a number of different "species" of agent, described in respective sections. Crucially, we did not change the core architecture or algorithm of the ToMnet to match the structure of the species, only the ToMnet's capacity.

3. Experiments

3.1. Random agents

We tested the ToMnet observer on a simple but illustrative toy problem. We created a number of different *species* of random agents, sampled agents from them, and generated behavioural traces on a distribution of random gridworlds (e.g. Fig A1a). Each agent had a stochastic policy defined by a fixed vector of action probabilities $\pi_i(\cdot) = \boldsymbol{\pi_i}$. We defined different species based on how sparse its agents' policies were: within a species $\mathcal{S}(\alpha)$, each $\boldsymbol{\pi_i}$ was drawn from a Dirichlet distribution with concentration parameter α. Species with small (large) α yield agents with deterministic (stochastic) policies. We then trained different ToMnet observers each on a single species of agent (details in Appendix A.2). We omitted the mental net for this task.

When trained on a species $\mathcal{S}(\alpha)$, the ToMnet learns to approximate Bayes-optimal, online inference about agents' policies. Fig 2a shows how the ToMnet's action likelihoods increase with more past observations, and how training the ToMnet on species with lower α yields priors that the policies are indeed sparser. We can also see how the ToMnet specialises by testing it on agents from different species (Fig 2b): the ToMnet makes better predictions about novel agents drawn from the species which it was trained on. Moreover, the ToMnet easily learns how to predict behaviour from mixtures of species (Fig 2d): when trained jointly on species with highly deterministic ($\alpha = 0.01$) and stochastic ($\alpha = 3$) policies, it implicitly learns to expect this bimodality in the policy distribution, and specialises its inference accordingly. We note that it is not learning about two *agents*, but rather two *species* of agents, which each span a spectrum of individual parameters.

Finally, the ToMnet exposes an agent embedding space; here it segregates agents along canonical directions by their 5-dim empirical action counts (Fig 2c).

In summary, without any changes to its architecture, a ToMnet learns a *general theory of mind* that is specialised for the distribution of agents it encounters, and estimates an *agent-specific theory of mind* online for each individual

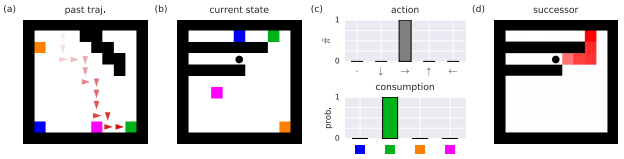

Figure 3. **ToMnet on goal-driven agents. (a)** Past trajectory of an example agent. Coloured squares: the four objects. Red arrows: agent's position and action. **(b)** Example query: a state from a new MDP. Black dot: agent position. **(c)-(d)** ToMnet's predictions for the query in (b), given the past observation in (a). SR in (d) for discount $\gamma = 0.9$. Darker shading: higher SR.

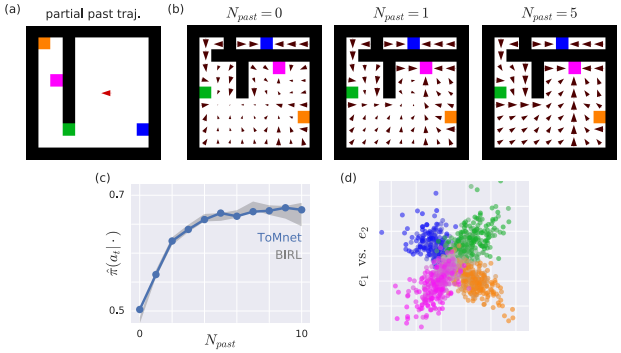

Figure 4. **ToMnet on goal-driven agents, continued. (a)** This ToMnet sees only snapshots of single observation/action pairs (red arrow) from a variable number of past episodes (one shown here). **(b)** Predicted policy for different initial agent locations in a query MDP. Arrows: resultant vectors for the predicted policies, i.e. $\sum_k \mathbf{a_k} \cdot \hat{\pi}(\mathbf{a_k}|x, e_{\text{char}})$. When $N_{\text{past}} = 0$, the predicted policy exhibits no net object preference. When $N_{\text{past}} > 0$, the ToMnet infers a preference for the pink object. When the agent is stuck in the top right chamber, the ToMnet predicts that it will always consume the blue object, as this terminates the episode as soon as possible, avoiding a costly penalty. **(c)** The average posterior probability assigned to the true action (16k sampled agents). Even when $N_{\text{past}} = 0$, this is greater than chance, since all agents in the species have similar policies in some regions of the state space. Model accuracy of BIRL shown as mean $\pm$ SEM over 16k sampled agents. **(d)** ToMnet's 2D embedding space (e_{char}). Colour: agents' ground-truth preferred objects; saturation increases with N_{past}.

agent that captures the sufficient statistics of its behaviour.

3.2. Inferring goal-directed behaviour

An elementary component of humans' theory of other agents is an assumption that their behaviour is *goal-directed* (Gergely et al., 1995; Woodward, 1998; Buresh & Woodward, 2007). We show here how the ToMnet observer learns how to infer the goals of reward-seeking agents.

We defined species of agents who acted within the gridworlds with full observability (Fig 3a). Each agent, $\mathcal{A}_i$, had a unique, fixed reward function over the four objects, and planned its behaviour through value iteration. We then trained the ToMnet to observe behaviour of these agents in randomly-sampled "past" MDPs, and to use this to predict the agents' behaviour in a "current" MDP. We detail three experiments below that explore the range of capabilities of the ToMnet in this domain.

First, we provided the ToMnet with a full trajectory of an agent on a single past MDP (Fig 3a). We then queried the ToMnet with the initial state of a current MDP (Fig 3b) and asked for a set of predictions: the next action the agent would take (Fig 3c top), whether it would consume each object (Fig 3c bottom), and a set of statistics about the agent's trajectory in the current MDP, the successor representation (SR; the expected discounted state occupancy; Dayan, 1993, Fig 3). The ToMnet's predictions qualitatively matched the agents' true behaviours.

Second, as a more challenging task, we trained a ToMnet to observe only partial trajectories of the agent's past behaviour. We conditioned the ToMnet on single observation-action pairs from a small number of past MDPs ($N_{\text{past}} \sim \mathcal{U}\{0, 10\}$; e.g. Fig 4a). In the absence of any past observations, the ToMnet had a strong prior for the behaviour that would be expected of any agent within the species (Fig 4b). With more past observations of an agent, the ToMnet's predictions improved (Fig 4c), yielding results comparable to Bayesian Inverse RL (BIRL).

Whether performance will scale to more complex problems is an open question. Since the ToMnet doesn't prespecify the generative model of behaviour, a bias/variance-tradeoff argument applies: the ToMnet will have higher sample complexity than BIRL when agents truly are noisy-rational utility maximisers; but unlike BIRL, it can flexibly deal with richer agent populations (Sections 3.3–3.5).

Unlike inverse RL, the ToMnet is also not constrained to explicitly infer agents' reward functions. Nevertheless, 2D character embeddings render this information immediately legible (Fig 4d), even when the ToMnet does not have to predict object consumption.

Finally, we enriched the agent species by applying a very high move cost (0.5) to 20% of the agents; these agents therefore generally sought the closest object. We trained a ToMnet to observe $N_{\text{past}} \sim \mathcal{U}\{0, 5\}$ full trajectories of randomly-selected agents before making its behavioural prediction. The ToMnet learned to infer from even a single trajectory which subspecies of agent it was observing, and predict future behaviour accordingly (Fig 5). This inference resembles the ability of children to jointly reason about agents' costs and rewards when observing short traces of past behaviour (Jara-Ettinger et al., 2016).

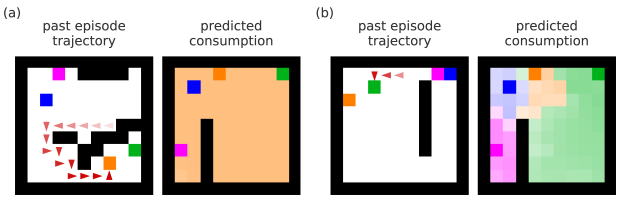

Figure 5. **ToMnet on greedy agents.** Left: a single past trajectory. Right: ToMnet predictions on a query MDP. Light shaded regions: the most probable object the agent will eventually consume, given that the agent is currently in that location. **(a)** After seeing the agent take a long path to the orange object, the ToMnet predicts it will try to consume the orange object on the query MDP, no matter its current location. **(b)** After seeing the agent take the shortest path to the green object, the ToMnet predicts it will generally consume a nearby object on the query MDP.

3.3. Learning to model deep RL agents

We next considered the ToMnet's ability to learn models for a richer population of agents: those with partial observability and neural network-based policies, trained using deep RL. This domain begins to capture the complexity of reasoning about real-world agents. So long as the deep RL agents share some overlap in their tasks, structure, and learning algorithms, we expect that they should exhibit some shared behavioural patterns. Individual agents may also exhibit idiosyncratic behaviour. There are thus opportunities to learn rich general and agent-specific theories of mind for such populations. Moreover, as tasks and agents become more complex, hand-crafting a Machine Theory of Mind to parse behaviour (e.g. Baker et al., 2011; Nakahashi et al., 2016; Baker et al., 2017; Lake et al., 2017) becomes increasingly intractable; instead we seek machines which learn to model others' minds autonomously (Botvinick et al., 2017).

We trained three different species of agents on gridworlds that included a subgoal object. Agents received maximum reward for reaching this subgoal first, then consuming a preferred object that differed from agent to agent. Consuming any of the non-subgoal objects terminated the episode. All agents used the UNREAL architecture (Jaderberg et al., 2017, Appendix D). One species of agent ("blind") was unable to observe the maze state at all, and could only observe its previous action (a_{t-1}) and reward (r_{t-1}), which it could integrate over time through its LSTM state. The second species had partial observability ("sighted"), but was stateless: these agents could observe the gridworld within a 5×5 window centred at their current location; their policies however were purely reactive, implemented via feedforward networks without any memory. The third species was both sighted (with partial observability) and stateful (with an LSTM-based policy). The ToMnet observed these agents' behaviour with *full* observability of the POMDP state. We trained the ToMnet on rollouts from 120 trained agents (3 species $\times$ 4 preferred objects $\times$ 10 initial random seeds). We held out a test set of a further 120 trained agents (i.e. 10 additional random seeds) for the results below.

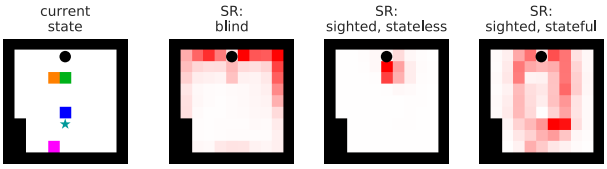

Figure 6. **Characterising trained neural-net agents.** ToMnet's prediction of agents' SRs given a query POMDP state at time $t = 0$ (left), as per Fig 3d. Star: the subgoal. Predictions made after observing behaviour on $N_{\text{past}} = 5$ past POMDPs from a sampled agent of each subspecies (always preferring the pink object).

Unlike previous experiments, these agents' behaviour depended on both their individual characteristics and their state; the ToMnet thus needed both a character net and a mental net to make the best predictions (Fig A2).

The ToMnet developed general models for the three different species of agents in its world. Fig 6 shows the ToMnet's predictions of SRs for the same query state, but given different past observations. Without being given the species label, the ToMnet implicitly infers it, and maps out where the agent will go next: blind agents continue until they hit a wall, then turn; sighted but stateless agents consume objects opportunistically; sighted, stateful agents explore the interior and seek out the subgoal.

With the higher dimensionality required to train the ToMnet on this task ($\mathbb{R}^8$), the embedding space lacked any discernable structure. This was likely due to the relatively deep prediction network, and the lack of explicit pressure to compress or disentangle the embeddings. However, the results were dramatically different when we added a variational information bottleneck to this layer (Alemi et al., 2016, Appendix A.4). By replacing the character embedding vectors e_{char} with Gaussian posteriors, $q(e_{\text{char}}|\cdot)$, the ToMnet was driven to disentangle the factors of variation in agent personality space (Fig 7). Moreover, the ToMnet even discovered unexpected substructure amongst the sighted/stateless subspecies, as it clustered sighted/stateless test agents into two subcategories (Fig 8a-b). Contrasting the ToMnet's predictions for these two clusters reveals the structure: each sighted/stateless agent explores its world using one of two classic memoryless wall-following algorithms, the *right-hand rule* or the *left-hand rule* (Fig 8c).

3.4. Acting based on false beliefs

Humans recognise that other agents do not base their decisions directly on the state of the world, but rather on an *internal representation*, or belief about the state of the world (Leslie, 1987; Gopnik & Astington, 1988; Wellman, 1992). These beliefs can be wrong. An understanding that others

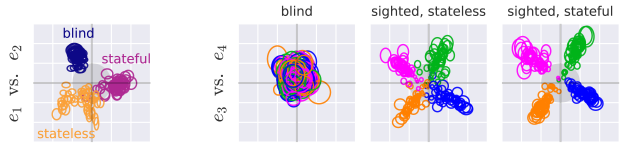

Figure 7. **Variational character embeddings.** Left: first two non-degenerate dimensions of $e_{\text{char}} \in \mathbb{R}^8$. Ellipses: Gaussian covariance (one stdev) of the posteriors $q(e_{\text{char}}|\cdot)$, coloured by agents' ground-truth species. Right: second two dimensions. Posteriors coloured by agents' ground-truth preferred objects. The ToMnet uses the first two dimensions to represent the agent's species, and the next two dimensions to represent its preferred object. When the agent is blind, the ToMnet represents the agent's preferred object by the prior, a unit Gaussian. All posteriors collapsed to the prior in the remaining four dimensions.

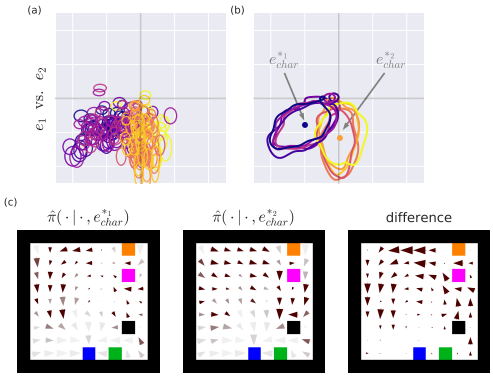

Figure 8. **Discovering subspecies.** **(a)** Posteriors, $q(e_{\text{char}})$, for sighted/stateless agents. Axes show the first two non-degenerate dimensions (as in Fig 7a). Each colour shows the posteriors inferred from a single deep RL agent from the test set, using different behavioural traces. **(b)** Marginal posteriors for the individual agents in (a), shown as iso-density contours, enclosing 80% of the total density. Dots: cluster means. **(c)** Predicted policy differences between agents in the two clusters in a query POMDP. Each panel shows predicted policy for different agent locations, as in Fig 4c. Left: ToMnet's prediction for an agent with e_{char} at the one cluster mean. Middle: at the other cluster mean. Arrows are darker where the two policies differ (higher D_{JS}). Right: vector difference between left and middle.

can have *false beliefs* has become the most celebrated indicator of a rich Theory of Mind (Baron-Cohen et al., 1985; Krupenye et al., 2016; Baillargeon et al., 2016).

Could the ToMnet learn that agents may hold false beliefs? To answer this, we needed a set of POMDPs in which agents could indeed hold incorrect information (and act upon this). We therefore 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 (Fig 9a-b). These *swap events* could only affect the agent when the objects' positions were within

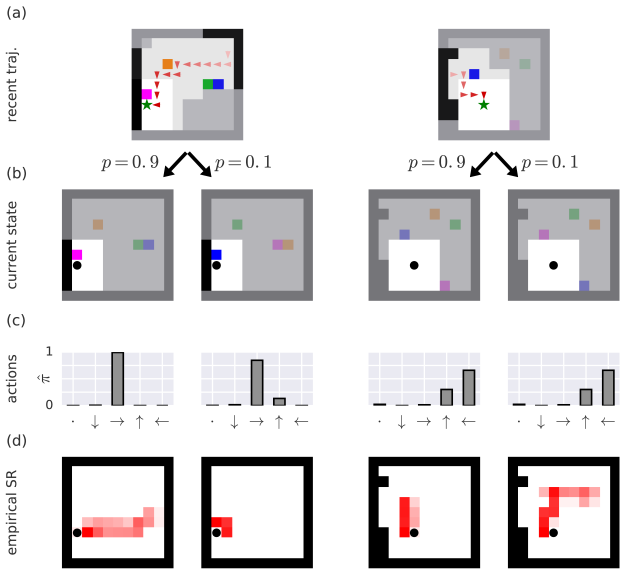

Figure 9. **Subgoal task, where agents can have false beliefs. (a)** Trajectory of an agent (red arrows) as it seeks the subgoal (star). Agent has partial observability: dark grey areas have not been observed; light grey areas have been seen previously, but are not observable at the time of subgoal consumption. **(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.

the the agent's current field of view; when the swaps occurred entirely outside its field of view, its policy at the next time step remained unaffected (Fig 9c, right), a signature of a false belief. As agents were trained to expect these low-probability swap events, they produced corrective behaviour as their policy was rolled out over time (Fig 9d, right). While the trained agents were competent at the task, they were not optimal.

In turn, we trained the ToMnet to predict the behaviour of these agents. We initially focused on agents with 5×5 fields of view. We trained the ToMnet on rollouts from 40 sighted/stateful agents (4 preferred objects $\times$ 10 random states), and tested it on a set of 40 held-out agents.

Our goal was to determine whether the ToMnet would learn a general theory of mind that included an element of false beliefs. However, the ToMnet, as described, does not have the capacity to explicitly report agents' (latent) belief states, only the ability to report predictions about the agents' overt behaviour. To proceed, we created a variant of the "Sally-Anne test", used to probe human and animal Theory of Mind (Wimmer & Perner, 1983; Baron-Cohen et al., 1985; Call & Tomasello, 2008). In this classic test, the observer watches an agent leave a desired object in one location, only for it to be moved, unseen by the agent. The

Figure 10. **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 ($\Delta\pi_L = (\pi(a_L|\text{no swap}) - \pi(a_L|\text{swap}))/\pi(a_L|\text{no swap}) \times 100\%$). If the agent can see that the object has moved from this location (swap dist ≤ 2), it will not return left. If it cannot see this location, its policy will not change. Right: ToMnet's prediction.

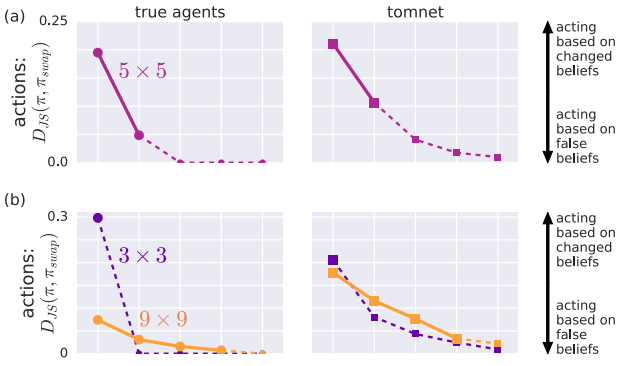

Figure 11. **Natural Sally-Anne test, using swap events within the distribution of POMDPs. (a)** Left: effect of swap events on 5×5 agents' next-step policies. Right: ToMnet predictions. **(b)** Results for a ToMnet trained on a range of agents with different fields of view. Showing only 3×3 and 9×9 results for clarity. Results for SRs shown in Fig A3. For a discussion about 3×3 agents' sensitivity to adjacent swap events: Appendix F.1.

subject, who sees all, is asked where the agent will seek the object. While infants and apes cannot explicitly report inferences about others' mental states, experimenters have nevertheless been able to measure these subjects' predictions of where the agents will actually go (Krupenye et al., 2016; Baillargeon et al., 2016).

We used the swap events to construct a gridworld Sally-Anne test. We hand-crafted scenarios where an agent would see its preferred blue object in one location, but would have to leave to reach a subgoal before returning to consume it (Fig 10a). During this time, the preferred object might be moved by a swap event, and the agent may or may not see this occur, depending on how far away the subgoal was. We forced the agents along this trajectory (off-policy), and measured how a swap event affected the agent's probability of moving back to the preferred object. As expected, when the swap occurred within the agent's field of view, the agent's likelihood of turning back dropped dramatically; when the swap occurred outside its field of view, the policy was unchanged (Fig 10b, left).

We presented these trajectories to the ToMnet (which had seen past behaviour indicating the agent's preferred object). Crucially, the ToMnet was able to observe the *entire* POMDP state, and thus was aware of swaps when the agent was not. To perform this task, the ToMnet needs to have implicitly learned to separate out what it *itself* knows, and what the agent can plausibly know. Indeed, the ToMnet predicted the correct behaviours (Fig 10b, right): when the world changes far away from an agent, that agent will pursue actions founded on false beliefs about the world.

We validated these results by looking at the ToMnet's predictions for how the agents responded to *all* swap events in the distribution of POMDPs. We sampled a set of test mazes, and rolled out the agents' policies until they consumed the subgoal, selecting only episodes where the agents had seen their preferred object along the way. At this point, we created a set of counterfactuals: either a swap event occurred, or it didn't. We measured the ground truth for how the swaps would affect the agent's policy, via the average Jensen-Shannon divergence (D_{JS}) between the agent's true action probabilities in the no-swap and swap conditions[1]. As before, the agent's policy often changed when a swap was in view (for these agents, within a 2 block radius), but wouldn't change when the swap was not observable (Fig 11a, left).

The ToMnet learned that the agents' policies were indeed more sensitive to local changes in the POMDP state, but were relatively invariant to changes that occurred out of sight (Fig 11a, right). The ToMnet did not, however, learn a hard observability boundary, and was more liberal in predicting that far-off changes could affect agent policy. The ToMnet also correctly predicted that the swaps would induce corrective behaviour over longer time periods, even when they were not initially visible (Fig A3a).

These patterns were even more pronounced when we trained the ToMnet on mixed populations of agents with different fields of view. Here, the ToMnet had to infer what each agent could see (from past behaviour alone) in order to predict their behaviour in the future. The ToMnet's predictions reveal an implicit grasp of how different agents' sensory abilities render them differentially vulnerable to

[1]For a discussion of the D_{JS} measure, see Appendix F.2.

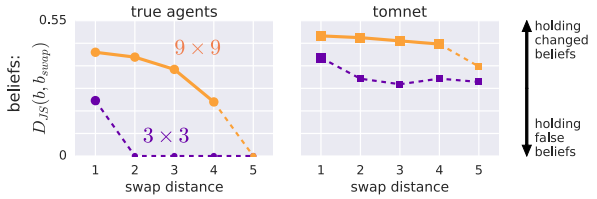

Figure 12. **Natural Sally-Anne task for reported beliefs.** The ToMnet captures the longer range over which the 9×9 agents update their beliefs (again, inferring a soft observability boundary).

acquiring false beliefs (Figs 11b, A3). Most surprising of all, the ToMnet learned these statistics even if the ToMnet had never seen swap events during training (Fig A4).

On the one hand, the ToMnet learns a general theory of mind that incorporates an implicit understanding that agents act based on their own representations of the world, even if they are mistaken. Yet this cognitive ability is unlikely due to the ToMnet architecture itself, which is indeed very straightforward. Rather, this work demonstrates that representational Theory of Mind can arise simply by observing competent agents acting in POMDPs.

3.5. Explicitly inferring belief states

While the ToMnet learns that agents can act based on false beliefs, it cannot explicitly report what these agents know and don't know about the world. It is also difficult to resolve any beliefs that do not manifest immediately in overt behaviour. We therefore extended the ToMnet to make declarative statements about agents' beliefs.

We constructed a supervised dataset of belief states by training the UNREAL agents to report their beliefs about the locations of the four objects and the subgoal at every time step, alongside their policy. Example belief states for the query MDP states in Fig 9a-b are shown in Fig A5a. These reported beliefs are just a readout from the LSTM state, so are not *causal* to the agents' policy.

In turn, we trained a single ToMnet on agents with a range of different fields of view. This ToMnet had to observe only agents' *overt* behaviour, and use it to predict, for a given query state, an agent's policy, consumptions, SRs, and reported beliefs (i.e., what the agent would say about the world if it were asked; e.g. Fig A5b).

The ToMnet learns agent-specific theories of mind for the different subspecies that grasp the essential differences between their belief-forming tendencies: agents with less visibility of changes in their world are more likely to report false beliefs, and behave according to them too (Fig 12).

Last of all, we included an additional variational information bottleneck penalty, to encourage low-dimensional abstract embeddings of agent types. The character embeddings again separated along the factors of variation (field of view and preferred object; Fig A6). This also reveals the ToMnet's ability to distinguish different agents' visibility: blind and 3×3 agents are easily distinguishable, whereas there is little in past behaviour to separate 7×7 agents from 9×9 agents (or little benefit in making this distinction).

This particular construction of explicit belief inference will likely not scale in its current form as it requires access to others' latent belief states for supervision. Here we provide this to the ToMnet via a rich communication channel; in the real-world, this channel is likely much sparser. Humans do, however, have privileged access to some of our own mental states through meta-cognition (though this data may be biased and noisy). It is also intractable to predict others' beliefs about every aspect of the world simultaneously. This may require solutions such as forming abstract embeddings of others' belief states that can be queried.

4. Discussion

We have shown that, through meta-learning, the *ToMnet* learns a general model for agents it encounters, and how to construct an agent-specific model online while observing a new agent's behaviour. The ToMnet can flexibly learn such models for many different species of agents, whilst making few assumptions about the generative processes driving these agents' decision-making. The ToMnet can also discover abstractions within the space of behaviours.

The experiments we pursued here were simple, and designed to illustrate the core ideas of such a system. There is much work to do to scale the ToMnet. First, we have worked entirely within gridworlds; these results should be extended to richer domains, e.g. 3D visual environments. Second, we did not limit the observability of the observer itself. This is clearly an important challenge within real-world social interaction, and is another inference problem (Baker et al., 2017). Third, there are many other dimensions over which to characterise agents, such as whether they are prosocial or adversarial (Ullman et al., 2009), reactive or able to plan (Sutton & Barto, 1998). Potentially more interesting is the possibility of using the ToMnet to discover new structure in the behaviour of either natural or artificial populations. Fourth, a rich Theory of Mind is likely important for many multi-agent decision making tasks, which will require situating ToMnet-like systems inside artificial agents. We anticipate many other needs: to enrich the set of predictions a ToMnet must make; to improve data efficiency at training time; to introduce gentle inductive biases; and to consider how agents might draw flexibly from their own cognition to inform their models of others. Addressing these will be necessary for advancing a Machine Theory of Mind that learns the rich capabilities of responsible social beings.

Acknowledgements

We'd like to thank the many people who provided feedback on the research and the manuscript, including Marc Lanctot, Jessica Hamrick, Ari Morcos, Agnieszka Grabska-Barwinska, Avraham Ruderman, Pedro Ortega, Josh Merel, Doug Fritz, Nando de Freitas, Heather Roff, Kevin McKee, and Tina Zhu.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

Albrecht, S. V. and Stone, P. Autonomous agents modelling other agents: A comprehensive survey and open problems. *arXiv preprint arXiv:1709.08071*, 2017.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv:1612.00410*, 2016.

Baillargeon, R., Scott, R. M., and Bian, L. Psychological reasoning in infancy. *Annual Review of Psychology*, 67: 159–186, 2016.

Baker, C., Saxe, R., and Tenenbaum, J. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Cognitive Science Society*, volume 33, 2011.

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

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

Botvinick, M., Barrett, D. G. T., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., Lillicrap, T., Modayil, J., Mohamed, S., Rabinowitz, N. C., Rezende, D. J., Santoro, A., Schaul, T., Summerfield, C., Wayne, G., Weber, T., Wierstra, D., Legg, S., and Hassabis, D. Building Machines that Learn and Think for Themselves: Commentary on Lake et al., Behavioral and Brain Sciences, 2017. *arXiv:1711.08378*, November 2017.

Brown, G. W. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13 (1):374–376, 1951.

Buresh, J. S. and Woodward, A. L. Infants track action goals within and across agents. *Cognition*, 104(2):287–314, 2007.

Call, J. and Tomasello, M. Does the chimpanzee have a theory of mind? 30 years later. *Trends in cognitive sciences*, 12(5):187–192, 2008.

Camerer, C. *Behavioral game theory*. New Age International, 2010.

Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Dennett, D. C. *The intentional stance*. MIT press, 1973.

Duan, Y., Andrychowicz, M., Stadie, B., Ho, J., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W. One-shot imitation learning. *arXiv:1703.07326*, 2017.

Everett, R. Learning against non-stationary agents with opponent modelling & deep reinforcement learning. *NIPS workshop*, 2017.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

Frith, C. D. and Frith, U. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.

Gallese, V. and Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998.

Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S. Taking the intentional stance at 12 months of age. *Cognition*, 56 (2):165–193, 1995.

Gopnik, A. and Astington, J. W. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pp. 26–37, 1988.

Gopnik, A. and Wellman, H. M. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2): 145–171, 1992.

Gordon, R. M. Folk psychology as simulation. *Mind & Language*, 1(2):158–171, 1986.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. In *NIPS*, pp. 3909–3917, 2016.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, 24(8):1979–1987, 2013.

Hochreiter, S., Younger, A. S., and Conwell, P. R. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pp. 87–94. Springer, 2001.

Hula, A., Montague, P. R., and Dayan, P. Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS computational biology*, 11 (6):e1004254, 2015.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. In *ICLR*, 2017.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. The naïve utility calculus: computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *COGSCI*, 2016.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114, 2016.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Leslie, A. M. Pretense and representation: The origins of" theory of mind.". *Psychological review*, 94(4):412, 1987.

Nakahashi, R., Baker, C. L., and Tenenbaum, J. B. Modeling human understanding of complex intentional action with a bayesian nonparametric subgoal model. In *AAAI*, pp. 3754–3760, 2016.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *ICML*, pp. 663–670, 2000.

Nowak, M. A. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

Oliehoek, F. A. and Amato, C. *A concise introduction to decentralized POMDPs*. Springer, 2016.

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

Raileanu, R., Denton, E., Szlam, A., and Fergus, R. Modeling others using oneself in multi-agent reinforcement learning. *arXiv preprint arXiv:1802.09640*, 2018.

Schmidhuber, J., Zhao, J., and Wiering, M. Simple principles of metalearning. *Technical report IDSIA*, 69:1–23, 1996.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Thrun, S. and Pratt, L. Learning to learn: Introduction and overview. In *Learning to learn*, pp. 3–17. Springer, 1998.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., and Tenenbaum, J. B. Help or hinder: Bayesian models of social goal inference. In *NIPS*, pp. 1874–1882, 2009.

Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18(2): 77–95, 2002.

Wang, Z., Merel, J. S., Reed, S. E., de Freitas, N., Wayne, G., and Heess, N. Robust imitation of diverse behaviors. In *NIPS*, pp. 5326–5335, 2017.

Wellman, H. M. *The child's theory of mind.* The MIT Press, 1992.

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

Woodward, A. L. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998.

Yoshida, W., Dolan, R. J., and Friston, K. J. Game theory of mind. *PLoS computational biology*, 4(12):e1000254, 2008.