Theory of Mind as Intrinsic Motivation for Multi-Agent Reinforcement Learning

Ini Oguntola [1] **Joseph Campbell** [1] **Simon Stepputtis** [1] **Katia Sycara** [1]

Abstract

The ability to model the mental states of others is crucial to human social intelligence, and can offer similar benefits to artificial agents with respect to the social dynamics induced in multi-agent settings. We present a method of grounding semantically meaningful, human-interpretable beliefs within policies modeled by deep networks. We then consider the task of *2nd-order* belief prediction. We propose that ability of each agent to predict the beliefs of the other agents can be used as an intrinsic reward signal for multi-agent reinforcement learning. Finally, we present preliminary empirical results in a mixed cooperative-competitive environment.

1. Introduction

The ability to infer the mental states of oneself and others – beliefs, desires, intentions, preferences, etc – is known as *theory of mind* (ToM) (Baker et al., 2011). Humans naturally build rich internal models of others, and are able to use these inferences to predict the behavior of others, to condition their own behavior, and to forecast social interactions (Georgeff et al., 1999). Theory of mind has long been studied within cognitive science and psychology (Premack & Woodruff, 1978), a fundamental aspect of human social intelligence that has been shown to develop in early childhood. (Ensink & Mayes, 2010; Astington & Edward, 2010).

Traditionally, agent-modeling approaches within reinforcement learning (RL) and imitation learning largely ignore the idea of internal mental states, typically only focused on modeling external actions (He et al., 2016; Wen et al., 2019). However, there is a growing body of work in the machine learning literature aimed towards developing artificial agents that exhibit theory of mind (Baker et al., 2011; Rabinowitz et al., 2018; Jara-Ettinger, 2019; Fuchs et al., 2021). Even beyond simply providing a helpful inductive bias for modeling behavior, ToM reasoning has the potential to enable the discovery and correction of false beliefs or incomplete knowledge, facilitate efficient communication and coordination, and improve human-agent teaming (Zeng et al., 2020; Sclar et al., 2022; Oguntola et al., 2021).

The work of (Aru et al., 2023) highlights key challenges regarding the difficulty of evaluating current deep learning ToM approaches. In particular, from a human perspective we may solve a task using an already-developed internal theory of mind, whereas an artificial agent may be able to learn simpler decision rules or take advantage of spurious correlations as shortcuts, and it is difficult to determine whether ToM has actually been learnt.

Here we consider the reverse – rather than solving a task and hoping it induces a theory of mind, we instead explicitly learn a theory of mind over semantically grounded beliefs, and use this as a signal to solve the task. Our fundamental research question is the following: can modeling other agents' *beliefs* serve as an intrinsic reward signal to improve performance in multi-agent settings?

In this paper we develop an approach to explicitly grounding semantically meaningful beliefs within RL policies. We then propose the use of ToM reasoning over the beliefs of other agents as intrinsic motivation in multi-agent scenarios. We run experiments in a mixed cooperative-competitive environment and show preliminary results that suggest this approach may improve multi-agent performance, with respect to both coordination and deception.

The primary contributions of this paper are the following:

- We develop an information-theoretic residual variant to the concept bottleneck learning paradigm (Koh et al., 2020) based on mutual information minimization.
- We utilize this approach to model semantically-meaningful belief states within RL policies.
- We propose the prediction task of second-order prediction of these beliefs (i.e. ToM reasoning) as intrinsic motivation.
- We demonstrate preliminary results that demonstrate improved performance in a mixed cooperative-

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Ini Oguntola <ioguntol@andrew.cmu.edu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

competitive environment.

2. Related Work

2.1. Intrinsic Motivation in Deep RL

Intrinsic motivation in reinforcement learning refers to the use of an additional reward signal to encourage particular agent behaviors without direct feedback from the environment on the task.

In the single-agent setting, common approaches to intrinsic motivation include "curiosity" to encourage visiting novel states (Pathak et al., 2017) and "empowerment" to encourage diversity of reachable states (Mohamed & Jimenez Rezende, 2015).

Most of these approaches can also be extended to the multi-agent setting, but the introduction of multiple agents inherently creates an inter-agent dynamic that can be explored as well. (Jaques et al., 2019) proposed an intrinsic reward for "social influence" by rewarding agents for having high mutual information between their actions. (Wang et al., 2020) develop similar approaches that reward an agent for influencing the state transition dynamics and rewards of other agents.

In constrast, our intrinsic reward approach is predicated on influencing the internal beliefs of other agents, rather than directly influencing their external states or actions.

2.2. Theory of Mind in Multi-Agent RL

Although RL often implicitly involves theory of mind via agent modeling, recent approaches have also sought to model this directly (Rabinowitz et al., 2018).

Within multi-agent reinforcement learning there have been a variety of approaches inspired by ToM reasoning, modeling beliefs (Fuchs et al., 2021; Wang et al., 2022; Sclar et al., 2022) and intents (Qi & Zhu, 2018; Xu et al., 2019). Other inverse reinforcement learning methods approach ToM-like reasoning by conditioning the reward function on inferred latent characteristics (Tian et al., 2021; Wu et al., 2023). Most of these are aimed at improving coordination in cooperative multi-agent scenarios, particularly with regard to communication (Sclar et al., 2022; Wang et al., 2022).

2.3. Concept Learning

Concept learning, generally speaking, is an approach to interpretability for deep neural networks that involves enforcing structure on the latent space to represent grounded, semantically meaningful "concepts".

One such approach is concept whitening (Chen et al., 2020), in which an intermediate layer is inserted for orthogonal alignment of data in the latent space with predefined human-interpretable concept labels, with concepts provided via auxiliary datasets. The restriction with this method is the inherent assumption that all concepts are non-overlapping.

Concept bottleneck models are a similar approach developed an approach that consists of a concept extractor directly supervised on concept labels, and a predictor network that generates an output from these concepts (Koh et al., 2020). While more flexible than concept whitening in the sense that it can encode any set of concepts, it still makes the assumption that the provided set of concepts alone is expressive enough for the predictive task; performance suffers when this not the case.

Some approaches mitigate this by combining the concept predictions with a residual extracted from the input, they either impose additional constraints (e.g. orthogonality) on the combined output that may not hold (Zabounidis et al., 2023), or they do not provide a way to directly ensure the information encoded by the residual does not overlap with the concepts (Yuksekgonul et al., 2022), allowing the model to effectively ignore concepts in its decision making process.

While prior work has used these approaches in the context of imitation and reinforcement learning (Oguntola et al., 2021; Zabounidis et al., 2023), in this work we specifically examine concept learning as a way to approach the challenge of grounding semantically meaningful *mental states* within policies. We also develop a residual variant that directly encourages decorrelation between concepts and residual while avoiding the introduction of any restrictive assumptions.

3. Method

3.1. Modeling Beliefs via Concept Learning

In deep reinforcement learning, policies are typically black box models that directly map states to actions. Our approach follows the paradigm of concept learning (Yi et al., 2018; Chen et al., 2020; Koh et al., 2020; Yeh et al., 2020; Oguntola et al., 2021; Zabounidis et al., 2023), which involves inserting an intermediate *concept layer* which is designed to align with human-interpretable "concepts", typically via a supervised auxiliary loss. In our setting, these concepts are designed to model *beliefs* about the environment. For instance, in an environment with a door, one could model the belief over whether the door is locked as a binary concept $b_{locked} \in \{0, 1\}$.

$$L_{belief} = \begin{cases} \mathrm{MSE}(\mathbf{b}, \mathbf{b}') & \text{if continuous} \\ \mathrm{CE}(\mathbf{b}, \mathbf{b}') & \text{if discrete} \end{cases} \tag{1}$$

where $\mathbf{b}$ is the agent belief vector, $\mathbf{b}'$ is the ground truth, MSE is the mean-squared error, and CE is the cross entropy loss.

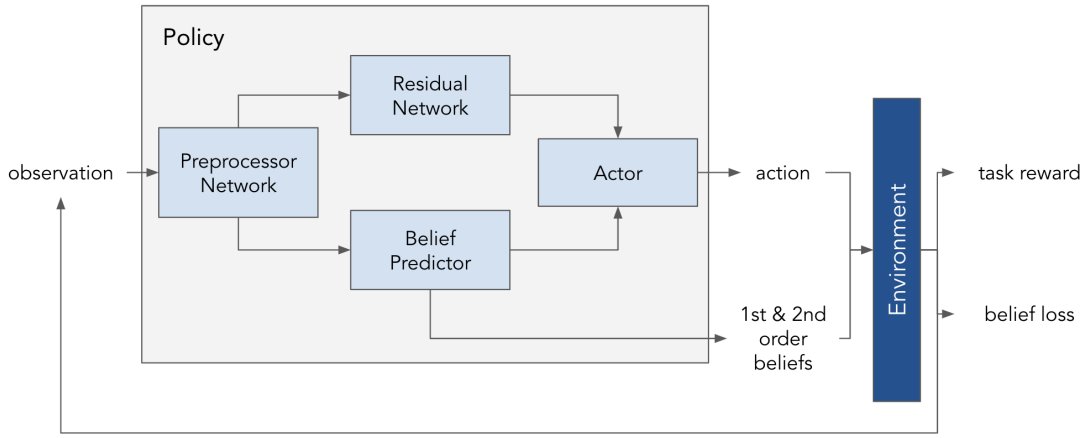

Figure 1. Policy models with 1st and 2nd-order belief prediction. The belief predictor is supervised by ground truth labels, and the residual network is regularized via mutual information minimization with respect to beliefs.

These beliefs are then used to generate an action. However, depending on the selection of beliefs, they alone may not be a sufficient signal to learn a policy that successfully solves a given task. We mitigate this by additionally introducing a *residual* – a compressed representation of the input that is concatenated to the belief vector. Given vector input $\mathbf{x}$, we have our residual network generate $r(\mathbf{x}) = \mathbf{z}$.

It is important that our residual and beliefs be disentangled – that is, the residual should not contain any information about the beliefs – as otherwise our model may simply learn to rely entirely on the residual and ignore the beliefs, which would compromise the interpretablity of the policy.

We approach "disentanglement" from a probability theory perspective, aiming to ensure that the belief and residual vectors are statistically independent. Here our goal is to minimize the mutual information between the belief vector and residual, which is zero if and only if they are independent. This measure can also be characterized as KL-divergence between the joint distribution and the product of the marginal distributions:

$$I(B;Z) = D_{KL}(\mathbb{P}_{BZ} \parallel \mathbb{P}_B \otimes \mathbb{P}_Z) \tag{2}$$

To achieve this, we utilize the variational approach from (Cheng et al., 2020) and minimize a contrastive log-ratio upper bound:

$$L_q(\theta) = -\mathbb{E}_{p_\sigma(\mathbf{b},\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{b})] \tag{3}$$

$$L_{residual}(\sigma) = \mathbb{E}_{p_\sigma(\mathbf{b},\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{b})] - \mathbb{E}_{p_\sigma(\mathbf{b})}\mathbb{E}_{p_\sigma(\mathbf{z})}[\log q_\theta(\mathbf{z}|\mathbf{b})] \tag{4}$$

where $\mathbf{b}$ is the belief vector, $\mathbf{z}$ is the residual vector, $p_\sigma(\mathbf{b},\mathbf{z})$ is the joint distribution of intermediate outputs from our policy, and $q_\theta(\mathbf{z}|\mathbf{b})$ is a variational approximation to the conditional distribution $p_\sigma(\mathbf{z}|\mathbf{b})$, modeled via a separate neural network trained to minimize negative log-likelihood $L_q(\theta) = -\log \mathcal{L}(\theta)$.

Unlike approaches based on concept whitening (Oguntola et al., 2021; Zabounidis et al., 2023), our method of disentanglement does not assume or impose any intra-dimensional orthogonality constraints within the concept (i.e. belief) or residual layers, but rather decorrelates the two vectors as a whole. Specifically, we make no restrictive assumptions that concepts are mutually exclusive, and also retain full multi-dimensional expressiveness within our residual representation while simultaneously minimizing correlation with our concept vector.

Finally, the concatenated output $(\mathbf{b},\mathbf{z})$ is fed into the rest of the actor network to generate an action. The concept layer and residual layer are trained by adding the additional loss terms to the objective function optimized by the reinforcement learning algorithm of choice. For our experiments we use the PPO objective from (Schulman et al., 2017), but generally speaking this approach is agnostic to the particular RL algorithm chosen.

$$L_{PPO}(\sigma) = \mathbb{E}_t[\min(r_t(\sigma)A_t, \text{clip}(r_t(\sigma), 1+\epsilon, 1+\epsilon)A_t)] \tag{5}$$

$$L_{policy} = \alpha L_{PPO} + \beta L_{belief} + \gamma L_{residual} \tag{6}$$

where $r_t(\sigma) = \frac{\pi_\sigma(a_t|s_t)}{\pi_{\sigma_{old}}(a_t|s_t)}$ is the PPO probability ratio, π_σ is the policy to be optimized, A_t is the advantage function, and $\alpha, \beta, \gamma, \epsilon > 0$ are hyperparameters.

During training, for each batch we optimize both the policy loss L_{policy} (with respect to the policy parameters σ) and the variational loss L_q (with respect to the variational parameters θ).

3.2. Second-Order Belief Prediction

In a multi-agent scenario where each agent is reasoning over the same set of beliefs over the environment, consider the *second-order belief* as one agent's prediction of another agent's beliefs. It is important to note that the first-order belief of an agent may be incorrect, in which case a correct second-order belief would successfully predict this false belief.

For instance, consider a scenario where a door is locked but agent A believes the door is unlocked. Agent B should ideally have 1) the first-order belief that the door is unlocked, and 2) the second-order belief that agent A thinks the door is locked.

Our approach proposes the use of second-order belief prediction as an intrinsic reward. Intuitively speaking, we want to incentivize each agent to 1) learn to predict the beliefs of other agents and 2) learn to behave in a way such that the beliefs of the other agents will be predictable (e.g. learning to observe other agents, learning to communicate, etc).

We do this by augmenting the agent's belief network to produce not only its own belief vector, but also a belief vector prediction for each of the other agents.

$$\mathbf{B} = [\mathbf{b} + f(\mathbf{x})_i]_{i=1}^{K} \tag{7}$$

where K is the total number of agents, $\mathbf{B}$ is the $K \times dim(\mathbf{b})$ second-order belief matrix, and $f : \mathbb{R}^{\dim(\mathbf{x})} \rightarrow \mathbb{R}^{K \times \dim(\mathbf{b})}$ is modeled by a neural network.

Rather than treat this as a directly-supervised auxiliary task, we instead include the second-order prediction loss as an additional reward term, as we want the policy's value estimation to be biased towards states where both the current and the **future** beliefs (or belief distributions) of the other agents tend to be predictable (e.g. states where it can gain information about other agents).

Then the intrinsic reward becomes the negative belief prediction loss:

$$r_{tom} = \begin{cases} -\frac{1}{K}\sum_{i=1}^{K} MSE(\mathbf{B}_i, \mathbf{b}^{(i)}) & \text{if continuous} \\ -\frac{1}{K}\sum_{i=1}^{K} CE(\mathbf{B}_i, \mathbf{b}^{(i)}) & \text{if discrete} \end{cases} \tag{8}$$

$$r = r_{task} + \lambda r_{tom} \tag{9}$$

where $\lambda \geq 0$ is a hyperparameter.

3.3. Training vs Execution

The training setup requires that all agents are trained in the manner previously described, and we assume that the beliefs of other agents are available during centralized training to calculate intrinsic reward.

During training we do not propagate gradients from the policy or reward through the 1st-order belief prediction network; that is, the 1st-order belief prediction network is only updated from the supervised belief loss on ground truth values from the environment, and is unaffected by the reward dynamics of the task. In combination with the mutual information regularization for the residual, this ensures that any belief information relevant to an agent's policy comes only from the agent's ability to infer the correct values of said beliefs from the environment. This approach eliminates any potential issues with a "malicious actor" purposefully generating incorrect belief predictions.

Execution, on the other hand, does not require beliefs or any inner states of other agents, and thus can be done with other policies that were not trained with our training setup or architecture – or even with human agents.

4. Experiments

4.1. ParticleWorld: Physical Deception

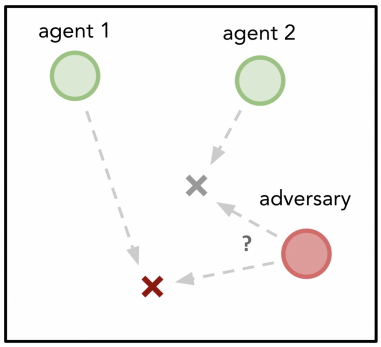

Figure 2. ParticleWorld physical deception environment.

We use a variant of the physical deception task described in (Lowe et al., 2017). This environment consists of N landmarks, N green "good" agents and a single red adversary agent within a 2D world.

In our variant, one of the landmarks is the "target", but neither the good agents nor the adversary are initially told which one. The N green agents receive a joint reward based on the minimum distance to the target landmark, with each agent's contribution weighted by a randomly generated reward coefficient $\eta_i \sim \text{Uniform}[0, 1]$. Similarly, the adversary is penalized based on its distance from the target.

The episode ends either after a fixed time-limit, or when the adversary reaches any landmark. If this is the target landmark, the adversary receives a positive reward, otherwise a

Table 1. Performance on ParticleWorld physical deception task, in various configurations. Here we present the mean cumulative reward of the final trained policies, averaged across 5 random seeds, where (good) is used to indicate the green good agents, and (adv.) is used to indicate the red adversary. With respect to beliefs, we vary whether the each policy generates 1st-order predictions, 2nd-order predictions, or none at all. Episode reward variance is given in parentheses.

1st-Order (Good)	1st-Order (Adv.)	2nd-Order (Good)	2nd-Order (Adv.)	Episode Reward (Good)	Episode Reward (Adv.)
No	No	No	No	1.889 ($\pm$ 0.23)	-15.32 ($\pm$ 0.51)
Yes	Yes	No	No	2.209 ($\pm$ 0.11)	-15.17 ($\pm$ 0.29)
Yes	Yes	**Yes**	No	**2.760 ($\pm$ 0.44)**	-17.78 ($\pm$ 0.32)
Yes	Yes	No	**Yes**	1.636 ($\pm$ 0.41)	**-14.01 ($\pm$ 0.30)**

negative penalty (both time–scaled).

$$r_{good}(t) = -\min_i \{d(\mathbf{x}_{i,t}, \mathbf{x}_{target})\} + d(\mathbf{x}_{adv}, \mathbf{x}_{target}) \quad (10)$$

$$r_{adv}(t) = -d(\mathbf{x}_{adv}, \mathbf{x}_{target}) + \mathbb{I}[\mathbf{x}_{adv} = \mathbf{x}_{other}](1 - t/T) - \mathbb{I}[\mathbf{x}_{adv} = \mathbf{x}_{target}](1 - t/T) \quad (11)$$

where d is Euclidean distance, $\mathbf{x}_{target}$ is the position of the target, $\mathbf{x}_{other}$ is the position of the non-target landmark, $\mathbf{x}_{i,t}$ is the position of good agent i at time t, $\mathbf{x}_{adv,t}$ is the position of the adversary agent at time t, and T is the maximum episode length.

The adversary is incentivized to find and navigate to the target as quickly as possible. On the other hand, the green agents are incentivized to keep the adversary uncertain as long as possible while accumulating reward.

Observations Each agent policy takes in a vector observation indicating the relative positions of landmarks and other agents. The good agents also can observe the weighted sum of their distances to the target landmark (weighted via their reward coefficients), whereas the adversary must rely on observing other agents' behavior to try and determine which landmark is the target.

Actions Each agent moves via a discrete action space.

Beliefs In this scenario each agent is trained with two sets of first-order beliefs:

1. Which landmark is the target?
2. What are the reward coefficients for each agent?

4.2. Training

We use Multi-Agent Proximal Policy Optimization (MAPPO) to train all agents in our experiments, under the paradigm of centralized training with decentralized execution (CTDE) (Yu et al., 2022). Our training procedure alternates between optimizing the policy for the good agents and the policy for the adversary, where one policy remains fixed and the weights other are trained; we swap every 100k timesteps.

5. Preliminary Results

We trained agents with various belief-prediction configurations on the physical deception task with $N = 2$ landmarks; training curves are shown in Figure 3, and the mean episodic reward achieved by the final policies are shown in Table 1. We report the mean episode reward obtained with the best hyperparameter setting over 20 episodes, for each of 5 random seeds.

We find that agents with the 2nd-order intrinsic reward perform significantly better in relation to the opposition. This phenomenon is observed for both the green good agents and the red adversary.

5.1. Qualitative Analysis of Observed Strategies

We qualitatively assess and summarize the strategies observed with the final trained policies from each of the configurations we considered below.

Baseline (no beliefs) Each green agent drifts towards a unique landmark. Red adversary appears to drifts randomly.

1st-order beliefs only (all agents) Similar behavior to baseline.

2nd-order beliefs (green agents) Each green agent drifts towards a specific landmark. In some episodes. green agents swap between landmarks.

2nd-order beliefs (red adversary) Red tends to be more decisive, moving quickly to landmark.

In both cases we observe that the incorporation of the 2nd-order intrinsic reward tends to lead to the exhibition of more complex strategies that do not seem to be discovered with

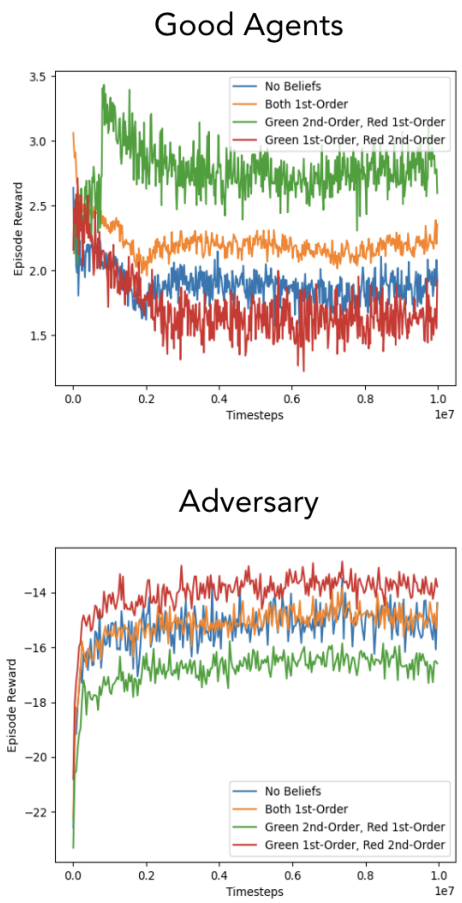

Figure 3. Training curves for agents on ParticleWorld physical deception task, in various belief configurations, averaged across 5 random seeds. Good agents trained with the 2nd-order belief intrinsic reward significantly outperform the other variants. Similarly, adversaries trained with the 2nd-order belief intrinsic perform better than their 1st-order belief and no-belief baseline counterparts.

the baseline MARL approach, or even when learning with 1st-order beliefs alone.

6. Ongoing and Future Work

Although preliminary results indicate our approach may be effective, they are with respect to a single, relatively simple environment. We are currently examining more complex multi-agent tasks with more varied social dynamics, and additionally scaling the approach to scenarios with more (or even an arbitrary number of) agents.

Beyond continuing to experiment with other environments, we are particularly interested in studying the efficacy of our approach in communication; both in more traditional cooperative scenarios as well as potentially in competitive tasks.

We are also interested in a more thorough investigation of our concept-residual approach in comparison with the standard whitening or bottleneck approaches (Chen et al., 2020; Koh et al., 2020).

References

Aru, J., Labash, A., Corcoll, O., and Vicente, R. Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review*, pp. 1–16, 2023.

Astington, J. W. and Edward, M. J. The development of theory of mind in early childhood. *Encyclopedia on early childhood development*, 14:1–7, 2010.

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

Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Ensink, K. and Mayes, L. C. The development of mentalisation in children from a theory of mind perspective. *Psychoanalytic Inquiry*, 30(4):301–337, 2010.

Fuchs, A., Walton, M., Chadwick, T., and Lange, D. Theory of mind for deep reinforcement learning in hanabi. *arXiv preprint arXiv:2101.09328*, 2021.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M. The belief-desire-intention model of

agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pp. 1–10. Springer, 1999.

He, H., Boyd-Graber, J., Kwok, K., and Daumé III, H. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*, pp. 1804–1813. PMLR, 2016.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Jara-Ettinger, J. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29: 105–110, 2019.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Information Processing Systems (NIPS)*, 2017.

Mohamed, S. and Jimenez Rezende, D. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

Oguntola, I., Hughes, D., and Sycara, K. Deep interpretable models of theory of mind. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 657–664. IEEE, 2021.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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

Qi, S. and Zhu, S.-C. Intent-aware multi-agent reinforcement learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 7533–7540. IEEE, 2018.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sclar, M., Neubig, G., and Bisk, Y. Symmetric machine theory of mind. In *International Conference on Machine Learning*, pp. 19450–19466. PMLR, 2022.

Tian, R., Tomizuka, M., and Sun, L. Learning human rewards by inferring their latent intelligence levels in multi-agent games: A theory-of-mind approach with application to driving data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4560–4567. IEEE, 2021.

Wang, T., Wang, J., Wu, Y., and Zhang, C. Influence-based multi-agent exploration. In *International Conference on Learning Representations*, 2020.

Wang, Y., Xu, J., Wang, Y., et al. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*, 2022.

Wen, Y., Yang, Y., Luo, R., Wang, J., and Pan, W. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Wu, H., Sequeira, P., and Pynadath, D. V. Multiagent inverse reinforcement learning via theory of mind reasoning. *arXiv preprint arXiv:2302.10238*, 2023.

Xu, K., Ratner, E., Dragan, A., Levine, S., and Finn, C. Learning a prior over intent via meta-inverse reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6952–6962. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/xu19d.html.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenenbaum, J. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Zabounidis, R., Campbell, J., Stepputtis, S., Hughes, D., and Sycara, K. P. Concept learning for interpretable multi-agent reinforcement learning. In *Conference on Robot Learning*, pp. 1828–1837. PMLR, 2023.

Zeng, Y., Zhao, Y., Zhang, T., Zhao, D., Zhao, F., and Lu, E. A brain-inspired model of theory of mind. *Frontiers in Neurorobotics*, 14:60, 2020.