Emergent deception and skepticism via theory of mind

Lion Schulz [*1] Nitay Alon [*1 2] Jeffrey S. Rosenschein [1 2] Peter Dayan [1 3]

Abstract

In complex situations involving communication, agents might attempt to mask their intentions, exploiting Shannon's theory of information as a theory of misinformation. Here, we introduce and analyze a simple multiagent reinforcement learning task where a buyer sends signals to a seller via its actions, and in which both agents are endowed with a recursive theory of mind. We show that this theory of mind, coupled with pure reward-maximization, gives rise to agents that selectively distort messages and become skeptical towards one another. Using information theory to analyze these interactions, we show how savvy buyers reduce mutual information between their preferences and actions, and how suspicious sellers learn to reinterpret or discard buyers' signals in a strategic manner. We discuss the implications of our findings for cognitive science and AI safety.

1. Introduction

Actions speak louder than words—sometimes enabling us to infer another person's beliefs and desires. Savvy speakers spin stories to fit their audience, like house buyers feigning disinterest to get a better deal. In turn, savvy listeners retaliate by ignoring them, like sellers sticking to their original prices. Here, we introduce a minimal two-agent task that captures the essence of such interactions and model agents' interaction using the reinforcement learning (RL) framework of Interactive Partially Observable Markov Decision Processes (IPOMDP; Gmytrasiewicz & Doshi, 2005), that endows agents with a theory of mind. We employ information theory to analyze the signalling behavior arising in this novel paradigm, showing how purely reward-maximizing agents endowed with a theory of mind (ToM) distort and re-interpret signals.

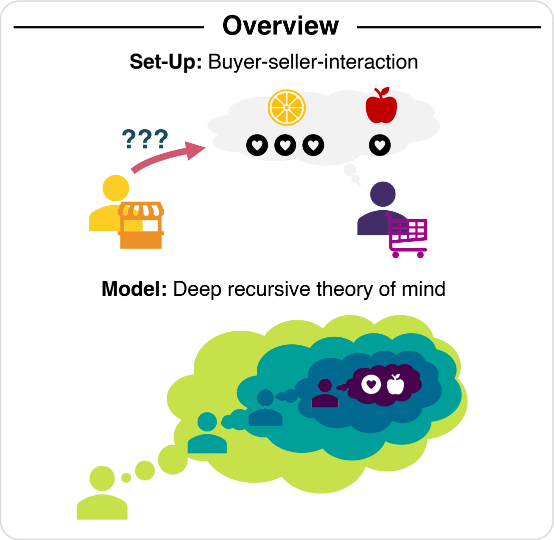
Overview
Set-Up: Buyer-seller-interaction
???
Model: Deep recursive theory of mind

Figure 1. To illustrate how theory of mind gives rise to deception and skepticism, we introduce a two-player game where a buyer and seller interact. We use IPOMDP to endow agents with recursive theory of mind.

To the AI community, our work demonstrates how complex signalling behavior can emerge via pure reward maximization (Silver et al., 2021). To the cognitive science and economics communities, we highlight computations that show how humans might optimally act when signalling using theory of mind. Information theory provides a tool with which to understand deception and skepticism. Our work also has broader implications, for example for dynamic pricing, privacy, recommender systems, and AI safety. We note that the results from the current work have been previously presented in a more condensed form (Alon et al., 2022).

1.1. Related work

In machine learning, inferring an agent's preferences, or utility function, is part and parcel of models that can be summarized as 'inverse reinforcement learning' (IRL) (Ng & Russell, 2000). These models observe an agent's actions and try to glean the its preferences from these observations.

[*] Equal contribution [1] Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tuebingen, Germany [2] Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel [3] Department of Computer Science, University of Tuebingen, Tuebingen, Germany. Correspondence to: Lion Schulz <lion.schulz@tue.mpg.de>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.

The core insight of these models is that agents only perform actions that are worth the cost. For example, if your colleague walks three blocks to a fancy cafe when there is a free coffee machine in the office, they likely award high utility to artisanal roasts. Algorithmic methods of exact or approximate Bayesian inference (Baker et al., 2011), including neural network distillations, can be used to carry out IRL (Rabinowitz et al., 2018; Oguntola et al., 2021).

In cognitive science, IRL has been used for models of the powerful inferences humans draw about one another (Baker et al., 2017; Jara-Ettinger, 2019). Chiefly among them, the 'Naive Utility Calculus' (Jara-Ettinger et al., 2016) proposes that humans reason in ways similar to Bayesian inference. This type of model has successfully explained how we reason about one another's kindness, knowledge, effort allocation, and skills (Xiang et al., 2023; Berke & Jara-Ettinger, 2021). Across fields, this ability to reason about others beliefs, desires and intentions has been referred to as "theory of mind" (ToM) (Premack & Woodruff, 1978), a trait hypothesized to be foundational to the human ability to perform complex social interactions (Ho et al., 2022; Ray et al., 2008; Xiang et al., 2012; De Martino et al., 2013; Rusch et al., 2020; Camerer et al., 2015; Hula et al., 2015; Zhao et al., 2013).

The 'Naive Utility Calculus', however, is not called 'naive' without reason, and its naivety extends to most other 'naive' IRL algorithms: they assume that the agent being observed is acting 'naively', that is, in a purely reward maximizing manner. However, agents are often aware that they are being watched. Such observer awareness (Miura & Zilberstein, 2021) matters less when the observer is an equally coffee-obsessed colleague, but is complicated when the observer can use its inferences to our detriment, for example, to increase the price of our favorite espresso—as, for example, in online dynamic pricing. Agents acting optimally should take such competitive scenarios into account when making their decisions—by inferring, and planning with, the inference of the observer in mind.

Such recursive reasoning about other agents extends theory of mind and has been used to explain a number of different phenomena in human interaction, from how we teach to how we lie and wage war (Ho et al., 2022; Crawford, 2003; Oey et al., 2022). Broadly, formal models of this recursivity (Doshi et al., 2020) tend to extend the simple inference in IRL and "plan through" this inference model. A particularly flexible instantiation of this is the mulitagent reinforcement learning framework Interactive Partially Observable Markov Decision Processes (IPOMDP; Gmytrasiewicz & Doshi, 2004). In essence, in an IPOMDP characterization, agents reason not only about uncertainty in the environment (as in a regular POMDP), but also about other agents' beliefs, desires, and intentions. Furthermore, agents do so in a recursively, at different levels of a 'cognitive hierarchy' (Camerer et al., 2004). That is, agents of different sophistication model agents with lesser sophistication ("I believe what you believe what I believe" and so on). We note that while IPOMDP is a very general model, others have also implement recursion in the MDP setting (Tejwani et al., 2022b;a).

In an IPOMDP, one agent's actions are interpreted by other agents as signals about latent characteristics that would be the target of IRL. Information theory (IT) is a particularly helpful tool to understand the messages that get sent among such sophisticated reasoning agents, particularly regarding deception (Kopp et al., 2018; Zaslavsky et al., 2020). This is because it allows us to measure the distortion of the signals transmitted between agents, and to pin down the deception and skepticism that possibly arise.

IT has also been used explicitly to model disinformation in multi-agent RL: Strouse et al. (2018) showed that training an agent to minimize or maximize the Mutual Information (MI) between its action and goals improves (mis-)communication in cooperative and competitive games. In contrast to Strouse et al. (2018), we show through simulations that theory of mind, when coupled with reward maximization and planning, suffices to give rise to such information-theoretic deception (and counter-deception, i.e., skepticism)—without the further need for handcrafting any value function. We examine in particular the extent to which signals about preferences become degraded and deciphered as theory of mind escalates.

2. Methods

2.1. Paradigm

As an example of the emergence of disinformation, we model a buyer and a seller interacting over three periods or stages (Fig. 2). Imagine a store owner offering items at stalls in various locations, and using observations of buyer behavior to set prices for subsequent sales to the buyer.

In the *first* stage, the buyer enters a simple T-Maze with two different items located at the ends of opposite arms (in our example, an apple and an orange). The buyer incurs travelling costs, denoted by the distances $d(\text{apple})$, $d(\text{orange})$, for travelling down one of the arms until it reaches and consumes one of the fruits. It then receives a reward based on its preferences. We denote the preferences over items by $r(\text{apple})$, $r(\text{orange})$, which are the rewards that the agent would receive from consuming them. Thus, the buyer's *immediate* utlity at stage one, $U_{B,1}$, from an item i_1 ($i_1 \in \{\text{apple}, \text{orange}\}$) at this stage is just the reward minus the cost incurred through the distance travelled:

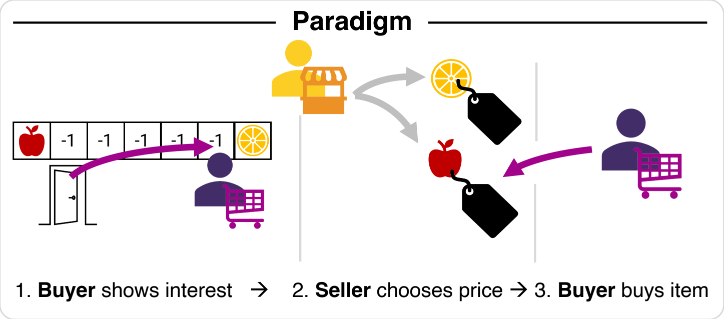

Figure 2. In our simulated paradigm, two agents interact, a buyer (orange) and a seller (purple). The buyer first chooses either of two differently rewarding objects, incurring cost on the way to the chosen object via the distance travelled. The seller observes this choice and the accompanying distances, draws inferences about the buyer's prices, and then prices the items accordingly. Finally, the buyer has to buy one of the items for the set price.

$$U_B^{t=1}(i_1, d) = r(i_1) - d(i_1) \tag{1}$$

Crucially, the buyer's decisions are observed by the seller (Miura & Zilberstein, 2021). Importantly, the seller is also aware of the buyer's travelling cost, but is *not* aware of its preferences towards either the apple or the orange, and, as we will see below, needs to infer them.

In the *second* stage, the seller uses their observation of the buyer to set prices, $m(i_3)$, for a future possible purchase of one of these two items. This process requires the seller to infer from the buyer's selection something about the buyer's preferences, so that the seller can set prices that maximize the seller's reward. As mentioned above, this requires a model of the buyer's behavior. We present these models, and describe the inverse inference process in a later section.

In the *third* stage, the buyer purchases one of the items for the set price, $m(i_3)$, and then consumes it, again receiving a reward $r(i_3)$ for this consumption according to its preferences:

$$U_B^{t=3}(i_3, m) = r(i_3) - m(i_3) \tag{2}$$

The overall (undiscounted) utility accumulated by each agent is thus as follows:

$$\begin{aligned} U_B^{total}(i_1, i_3, d, m) &= \\ U_{B,1}(i_1, d(i_1)) &+ U_{B,3}(i_3, m(i_3)) \end{aligned} \tag{3}$$

The seller's reward is just the price for the item the buyer buys:

$$U_S(i_3) = m(i_3) \tag{4}$$

Note that $d(i_1)$ is environmental (distance to items), while $m(i_3)$ is set by the seller. For illustrative purposes, we here restrict the preferences of oranges and apples to sum to 10. We impose the same restriction on the walking distances in the first stage and the prices that the seller can set in the second stage.

2.2. Model and agents

We study different levels of sophistication of buyers and sellers arising from the complexity of their ToM using the IPOMDP framework. IPOMDP augments the POMDP framework (Kaelbling et al., 1998) to include inference about other agents' (Fig. 2). This reasoning includes inferring about other agents' utility (IRL) as well as inferring about their beliefs. Since these beliefs include the other agent model of the inferring agent, ToM is defined recursively (see Fig. 3). Unlike Camerer et al. (Camerer et al., 2015), we assume a strict nesting, where each ToM-level uses only a model of the agent one step below on the ToM ladder as do Gmytrasiewicz & Doshi (2005).

As the turns in this game alternate, the ToM levels of buyer and seller also alternate (Hula et al., 2015), starting from the simplest buyer, which we denote as ToM($k = -1$). The **ToM**(-1) **buyer** independently decides based on distance/price and preferences at the first and third stages without taking into account the seller. That is, this naïve buyer solves what amounts to the same utility maximization problem at each step where it gets a choice. As a result, its action-values Q at a given state are simply defined as the aforementioned utilities.

$$Q_{k=-1}^{t=1}(i_1, d(i_1)) = U_B^{t=1} = r(i_1) - d(i_1) \tag{5a}$$

$$Q_{k=-1}^{t=3}(i_1, d(i_1)) = U_B^{t=1} = r(i_3) - m(i_3) \tag{5b}$$

After the buyer computes these Q-values, it then selects an item using a SoftMax policy (Equations 6a, 6b) with inverse temperature β. If we use j to denoting the 'other' item (j represents 'apple' if i represents 'orange'), then the probabilities of the actions a_1 and a_3 of selecting item i_1

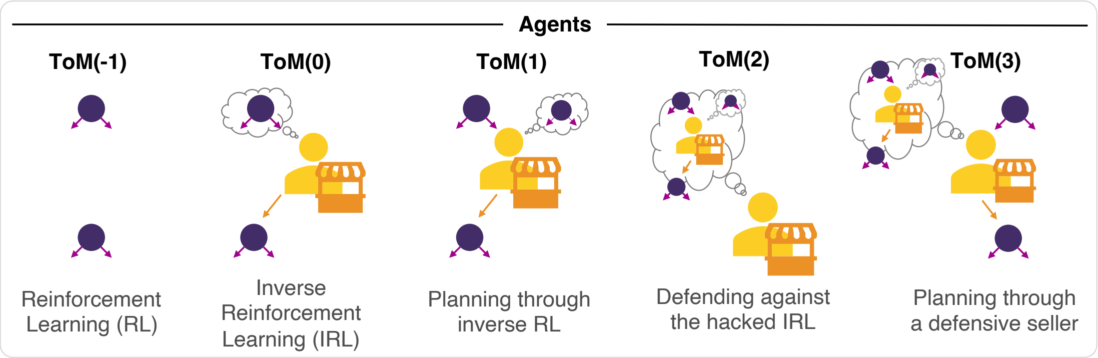
Agents
ToM(-1)
ToM(0)
ToM(1)
ToM(2)
ToM(3)
Reinforcement Learning (RL)
Inverse Reinforcement Learning (IRL)
Planning through inverse RL
Defending against the hacked IRL
Planning through a defensive seller

Figure 3. **Model:** We model agents of different levels of theory of mind (ToM). We begin with a simple reinforcement learning buyer (purple) that makes the first and second choice independently, and a regular inverse reinforcement learning seller (orange), ToM(1), who draws inferences about this buyer. The ToM(1) buyer plans through this seller's inference process, and the ToM(2) seller tries to defend itself against such a hack. The ToM(3) buyer finally tries to exploit the ToM(2) seller's inference process.

and i_3 at stages 1 and 3 are:

$$P^{t=1}_{k=-1}(a_1 = i_1|\mathbf{r}, \mathbf{d}) = \sigma(\beta\Delta^{t=1}_{k=-1}(i_1, j_1)) \tag{6a}$$
$$P^{t=3}_{k=-1}(a_3 = i_3|\mathbf{r}, \mathbf{m}) = \sigma(\beta\Delta^{t=3}_{k=-1}(i_3, j_3)) \tag{6b}$$
$$\text{with} \quad \Delta^t_k(i, j) = Q^t_k(i, d(i)) - Q^t_k(j, d(j))$$

In RL terms, these action probabilities are referred to as the agent's policy. Since $r(i) + r(j) = d(i) + d(j) = m(i) + m(j)1$, these expressions only depend on one member of the pairs of rewards, distances and prices.

To make inferences about the preferences of the ToM(-1) buyer, the **ToM**(0) **seller** performs Bayesian inverse reinforcement learning (Ramachandran & Amir, 2007) on the buyer's first stage actions and sets the prices accordingly. The IRL process (Eq. 7) inverts the selection using Eq. 6a and reweighed using the prior $p(r(i_1))$, assumed to be a uniform distribution $\mathcal{U}[0, 10]$ in this paper.

$$p_{k=0}(\mathbf{r}|a_1 = i_1) \propto \hat{P}_{k=-1}(a_1 = i_1|\mathbf{r}, \mathbf{d})p(\mathbf{r}) \tag{7}$$

Here, we note that what the seller assumes to be the buyer's likelihood, $\hat{P}_{k=-1}(a_1 = i_1|\mathbf{r}, \mathbf{d})$, in this case is just the ToM(-1)'s policy we defined above in equation 6a.

Given the posterior beliefs, the ToM(0) seller sets the optimal prices via expected utility maximization. That is, the optimal price of item i is set via:

$$m^*_{k=0}(i_3, a_1) = \underset{m(i)}{\text{argmax}} \{ E[U_S(i)|a_1]\} \tag{8}$$

We outline the full expansion of this equation in the appendix (Eq. 12). In brief, for every possible item price $m(i)$ the seller computes the subjective probability that the ToM(-1) buyer will buy the item at that said price: $\hat{P}_{k=-1}(a_3 = i|r(i), m(i))$. The potential revenue, $U_{S,total}(i_3) = m(i_3)$, is weighted by the posterior probability over the buyer's preferences computed via the inverse reinforcement learning spelled out in Equation 7. Note that while the buyer's decision-making is stochastic, the seller's policy sets a deterministic price for each item. While this has so far been a simple pairing of reinforcement learner and inverse reinforcement learner, agents with higher theory of mind will make more complex decisions.

Specifically, the **ToM**(1) **buyer** takes the ToM(0) price computation into account, and plans through it to optimize the sum of both first and third stage payoffs. Crucially, this model-based planning takes into account the seller's inverse reinforcement learning, as the expectation in Eq. 13 is with respect to this instance of IRL. This planning is computed through a full planning tree span. After the simulation terminates it outputs the action-values $(Q(a))$ for each item selection in the first phase. The ToM(1) selects the first action via a SoftMax policy like the ToM(-1) buyer. Since the buyer cannot affect the seller's behavior in the last step, the ToM(1) acts an identical way to the ToM(-1), that is, it selects an item via utility maximization.

The key factor in the ToM(1) policy is the belief manipulation of the ToM(0) seller. Since the pricing policy affects the overall utility of the buyer, but is affected by the buyer's decision in the first step, we can express the Q-values of the $k = 1$ buyer at the first stage as:

$$Q^{t=1}_{k=1}(a_1 = i_1, \mathbf{d}) = \tag{9}$$
$$U^{t=1}_B(i_1, \mathbf{d}) + E[U^{t=3}_B(i_3, m^*_{k=0}(i_3, a_1))|a_1 = i_1]$$

We present the full equation in the appendix. Unpacking it, the ToM(1) agent performs a thought experiment in which it envisions itself acting as the ToM(0) seller—observing the first item pick and setting the prices accordingly. Using its mental model of the seller it can simulate an inference process (Eq. 7) and the consequent prices (Eq. 8). Thus, the ToM(1) buyer can predict how its action in the first step affects the potential reward in the last phase through full mentalization of the seller's inference and learning process.

Crucially, while the ToM(0) seller makes inferences about the ToM(-1) buyer's utility from behavior, which is the equivalent of IRL, the ToM(1) buyer makes inferences about the optimal pricing of the ToM(0) seller given the buyer's own selection in the first phase. Note that in the last step of the task the prices are given. Hence, the seller selects an item that maximizes its utility similarly to the ToM(-1) item selection (6b). Hence, the strategic aspect of the buyer's planning is first-item selection, as it is this that affects the beliefs of the seller about the buyer's preferences.

During planning, the buyer 'imagines' its own initial action, and then simulates the seller's best response as in Eq. 8. This simulation performs the hypothetical IRL of the seller and its consequent beliefs, and then computes the optimal price from its policy. Thus, as discussed in the next section, the ToM(1) buyer improves its utility through shaping the beliefs of the ToM(0) seller. The ToM(1) choice at the first stage is selected through a SoftMax policy (similar to the ToM(-1) policy):

$$P^{t=1}_{k=1}(a_1 = i|\mathbf{r}, \mathbf{d}) = \sigma(\beta[Q^{t=1}_{k=1}(i_3, d(i_3)) - Q^{t=1}_{k=1}(j_3, d(j_3))]) \tag{10}$$

The **ToM**(2) **seller** models the buyer as a ToM(1) buyer, and is aware of the manipulation schema deployed by it. In essence, it uses the same principles as the ToM(0) seller for its inverse RL, but unlike the ToM(0) seller, it learns to treat the signal provided (item selection) with caution by using the ToM(1)'s policy as its likelihood in the Bayesian update:

$$p_{k=2}(\mathbf{r}|a = i_1) \propto \hat{P}_{k=1}(a = i_1|\mathbf{r}, \mathbf{d})p(\mathbf{r}) \tag{11}$$

Lastly, the **ToM**(3) **buyer** again attempts to 'hack' the ToM(2) beliefs in a similar manner to the way the ToM(0) attempts to manipulate the ToM(0) seller. This deception depends on the 'wiggle room' left, given the defensive policy adopted by the ToM(2) seller.

To hone in on an important point: each agent higher up in the cognitive hierarchy nests the inference and planning of those agents below it in the cognitive hierarchy. Essentially, there is ever more sophisticated reinforcement learning (planning) that gives rise to a policy, and ever more sophisticated inverse reinforcement learning that inverts this policy.

In the next section we present the results of the dyadic simulations when we compute the optimal policies of the agents.

2.3. Information Theory, Deception and Skepticism

As we briefly highlighted in the introduction, information theory presents an elegant tool to analyze the signals sent between agents, and their deceptive as well as skeptical nature. It allows us to do so from different perspectives:

1. From the perspective of the sender of a message, we can ask *how much a message reveals* about something it wants to hide. As we will see, this will be particularly relevant when asking how much the buyer's actions reveal about its preferences. Information theory allows us to capture this using the Mutual Information between the buyer's actions and its preferences, $I(\mathbf{r}, a_1)$.

2. From the perspective of the receiver of the message, we can analyse *how much credence is lent to a signal*. We can do so by calculating how much a receiver's beliefs change in response to a signal. Information Theory lets us do this via the Kullback-Leibler (KL) Divergence between a receiver's prior and its posterior once it has seen a message. Here, we are interested in this case for the seller before and after it has seen the buyer's action: $D_{KL}((p(r|a_1)||p(r))$. Essentially, the lower this divergence, the more skeptical an agent.

3. Finally, we can take a more bird's eye view of an interaction and ask *how much a given signal sent by a sender is misinterpreted by a receiver*. Again, we can do so using KL-Divergence, for example between what the receiver assumes is a sender's policy and what the sender's actual policy is. In our case, we are for example interested in how simpler sellers might be led astray by higher level theory of mind buyers. We measure this as the KL-distance between what a ToM(k) seller *assumes* to be the buyer's policy (i.e. the ToM$(k-1)$ buyer) and the ToM$(k+1)$'s actual policy, for example for the the case of the ToM(0) buyer: $D_{KL}((p^{t=1}_{k=1}(a_1|r, d)||p^{t=1}_{k=-1}(a_1|r, d))$. Essentially, this quantifies how effective a senders deception is, and is therefore different from the mutual information outlined above, which is more about the mere hiding of information (Kopp et al., 2018).

3. Results

We present the agents' policies resulting from this progressive, recursive modelling. As described above, the only strategic action of the buyer is its first move; hence we compare this action across different ToM levels. In addition, we present the seller's corresponding prices. We describe how each buyer's behaviour can be seen as a "best-response" to its perceived opponent. In turn, we discuss how the sellers respond to these policies. Throughout, we quantify these behavioural dynamics with key information theoretic metrics and highlight the relation between the two in light of the cognitive hierarchies. We describe the intricacies of these policies in substantial detail in order to exploit the simplicity and transparency of our setting.

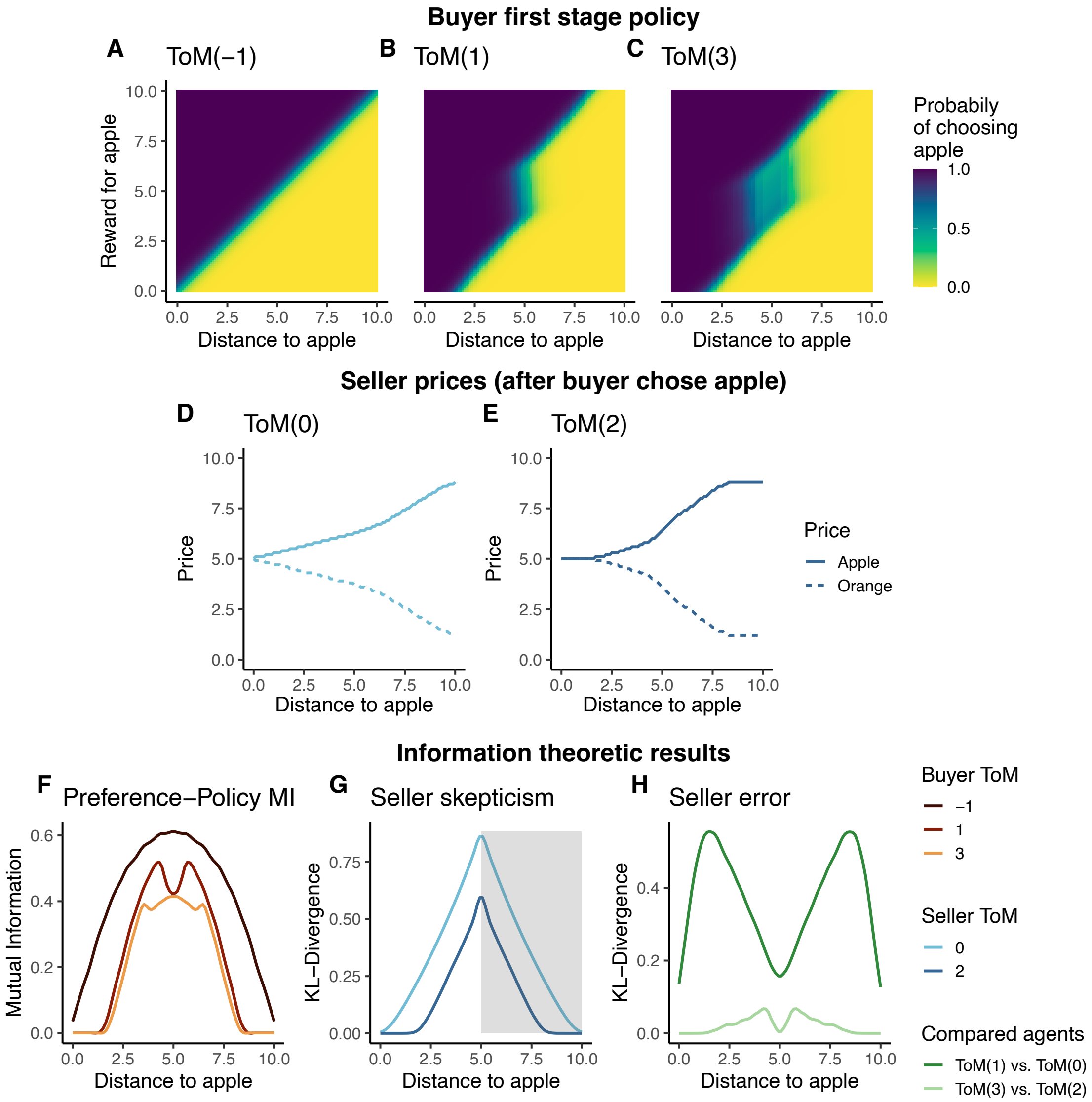

Figure 4. **(A-C)** Buyer policy in first stage as a function of the distance $d(\text{apple})$ and the preference towards the apple, $r(\text{apple})$, shown by different ToM-levels. Here, we are using a soft policy with a temperature of $\beta = .5$, as for example in equation 6a. The policies are symmetric because of the constraints. Notice the ruse in ToM(1) and ToM(3), who shift their policies. **(D-E)** Seller prices, $\mathbf{m}$, after a buyer chose the apple as a function of the distance to the apple, $d(\text{apple})$, for the two different seller ToM-levels. ToM(2) discards the evidence at the extremes. **(F)** Amount of deception by the different ToM buyers quantified by the mutual information between the buyer's apple preferences and the probability that they will pick an apple. ToM(1) and ToM(3) show lower values, effectively hiding their preferences. **(G)** Strength of the seller's belief update quantified by the KL-Divergence between their (flat) prior and posterior over the apple preferences, after observing the buyer choose the closer and thus more likely object (apple in left half, orange in shaded right half). **(H)** Dissimilarity between the $ToM(k)$ seller's assumed policy and the $ToM(k+1)$ buyer's actual policy, simultaneously showing the hacking success of the buyer and error of the seller.

We begin with the **ToM(-1) buyer**. As we outlined in the model section, this buyer acts naïvely, maximizing the utility of each stage separately. Fig. 4A shows the probabilities of choosing an apple, the x-axis describing the distance from the entrance to the apple, $d(\text{apple}_1)$, and the y-axis the reward derived from consuming the apple at the end of the corridor, $r(\text{apple}_1)$. The colours represent apple selection probability. Here, due to the symmetric nature of the problem we only discuss the apple selection.

The policy resulting from the ToM(-1)'s value computations are straightforward: The apple is more likely to be chosen when it is closer (left of x-axis) and more preferred (top of y-axis). This behaviour is well aligned with the lack of opponent model—the ToM(-1) buyer does not try to conceal its preferences, since it does not model the seller's pricing scheme and so believes that its behaviour has no consequences.

As we noted, we can re-express such a lack of concealment in information-theoretic terms by measuring how informative the buyer's choice is about its preferences, i.e., the Mutual Information (MI) between the buyer's preferences and their initial choice, $I(\mathbf{r}, a_1)$. This MI is shown in the darkest curve in Fig. 4F—the action is generally informative, particularly when the fruits are nearly equidistant from the buyer. Here, the buyer's choice reveals most about its preferences.

The ToM(-1) buyer in turn is the input to **the ToM**(0) **seller**. This seller can translate what it knows about the naive ToM(-1) policy into prices via simple inverse reinforcement learning. Fig. 4D shows these prices for the case that the buyer has chosen the apple in the first stage. Since the ToM(-1) signal is reliable, the distance covered by the buyer is a good proxy for its preferences. For example, if the apple is situated 8 steps from it, and the buyer picks it, then the seller can infer that the buyer's utility is at least 8 and set the price just shy of this. This all but guarantees that the apple will be selected at the last phase.

Using IT, we re-express this "trust" in the buyer's action in mathematical terms. Specifically, we measure the strength of the seller's belief update via the KL-divergence (KLD) between its (flat) prior and posterior, $D_{KL}((p(r|a_1)||p(r))$. In the lighter curve in Fig. 4G, we show the ToM(0) seller's KLD associated with the choice of the more likely item, which is apple when apple is closer and orange when orange is closer (the latter shown in the shaded area). This highlights how the seller uses every decision of the buyer as a signal, irregardless of the maze set-up.

Aiming to get the best price possible, the **ToM**(1) **buyer** attempts to hack this pricing scheme by playing what amounts to a gambit. This manipulation is manifested in three different ways that are evident in Fig. 4B. First, the ToM(1) buyer adopts a deceptive maneuver for the case when the apple is close but *undesired* (the uniformly dark-purple part between $d(\text{apple}) = 0$ and $d(\text{apple}) = 2$). In this setting, the buyer knows that the ToM(0) seller would interpret an orange selection as a signal for high orange preference. Hence it selects the apple despite its lower appeal to convince the seller that it prefers the apple to orange and gain a lower price for the orange down the road.

The second deception takes place in the shift of location of the indifference line (for example when $d(\text{apple}_1) = 7$). Here, the probability of selecting an apple rises above parity only when the preferences are about 7.5 instead of 7 (the expected behaviour by the observing ToM(0)). This "delayed" selection shifts the posterior beliefs of the ToM(0) seller to believe that the buyer prefers oranges more than it actually does and the buyer gets a discounted price for the apple.

Lastly, in the central region ($d(\text{apple}_1) = 5$) the ToM(0) policy outputs a slightly off-diagonal line of indifference. Again, from the perspective of the ToM(0) observer, the probability of selecting an apple in this setting, when the apple preference is 6.0, are almost 1.0 and not 0.5—thus the naïve ToM(0)'s inverse RL process is inaccurate.

We can express this ruse in information-theoretic terms in two ways. Returning to the MI between (naïve) preferences and policy, we show how the ToM(1) buyer manages to significantly reduce this informativeness about its preferences, particularly when one of the items is close (the medium line in Fig. 4F).

As we discussed, we can measure the success of the buyer's ruse by asking how wrong the ToM(0) seller's model is. We do so by measuring the KLD between what the ToM(0) seller *assumes* to be the buyer's policy (i.e., a ToM(-1) buyer) and the ToM(1)'s actual policy, $D_{KL}(p^{t=1}_{k=1}(a_1|r, c)||p^{t=1}_{k=-1}(a_1|r, c))$. This belief discrepancy is shown in a darker green line in Fig. 4H, highlighting the large discrepancies.

Having access to this ToM(1) buyer model, the **ToM**(2) **seller** becomes skeptical about the buyer's actions, and adjusts its pricing appropriately. We show this pricing in Fig. 4E. When the maze setting enables the ToM(1)'s bluff (for example, when the distance to the apple $d(\text{apple}_1) = 1$), observing an apple selection provides the seller with no information about the buyer's preferences (compare the previously discussed uniformly purple policy of the ToM(1) agent in Fig. 4B, and notice how it always chooses the apple regardless of preference). As a result, the seller ignores the distance travelled by the buyer and keeps pricing the items equivalently at these distances.

As the cost of bluffing increases, the seller adapts the apple price to match, but does so at a sub-linear rate, still remaining suspicious of the buyer's choice—which is warranted by the buyer's policy of over-selecting the under-preferred item. However, when the apple is farther away than the orange (right half of the plot), this logic switches, and picking the apple now becomes a very strong signal of the buyer actually liking the apple. This is because the ToM(1) buyer is now more likely to employ a similar ruse towards the orange, and would only pick an apple when it has a really strong preference towards it.

Information theory again lets us formalize this skepticism via the KL-divergence as a function of the item more likely to be picked (see the dark line in Fig. 4G, which shows how the ToM(2) seller's belief about the buyer is affected less, or, when the items are closer, not at all, by the buyer's actions).

The **ToM**(3) **buyer** attempts to manoeuvre around this skeptical pricing to achieve the best overall reward. However, it is essentially cornered and can only attempt minimal ruses in a few possible game settings, particularly when the items are roughly equidistant (see Fig. 4C). In fact, it must act like a ToM(1) buyer because the somewhat paranoid ToM(2) would otherwise overprice the preferred item heavily.

This inability to outmanoeuvre the seller significantly has an information-theoretic consequence. While the Mutual Information between policy and naïve preferences of the ToM(3) buyer is, in some regions, slightly lower than the ToM(1)'s, the ToM(3) buyer cannot mischaracterize its preferences further (lightest curve in Fig. 4F). Equally, the discrepancy between the ToM(2) seller's assumptions about the ToM(3) buyer and the truth is much less than that for the ToM(0) seller and ToM(1) buyer pair (light curve in Fig. 4H). Note that the ToM(2) dissimilarity increases in regions where the ToM(0) dissimilarity decreases, showing the ToM(3)'s attempts at deception.

4. Discussion and future work

4.1. Summary of results

Our work shows how purely reward-maximizing agents can appear to engage in complex signalling behavior, as captured by information theory. Crucially, unlike Strouse et al. (2018), we do this in the absence of any hand-crafted value function and only rely on theory of mind and planning. We show how, through a form of planning that is opponent-aware, agents can exploit other agents' inference processes. More specifically, we show how a sender can purposefully reduce the informativeness of their actions and target the inferences they expect a receiver to perform. On the other hand, we show how agents can defend themselves from this manipulation by (partially) ignoring the behaviour they observe. This counter-deception can be interpreted as skepticism and we illustrate it both in policies and in inference.

We quantified the extent of manipulation in these deceptive behaviours using information-theoretic metrics. Our results follow the conceptual ideas presented by (Kopp et al., 2018): agents learn to distort the communication by reducing informative information and deliberately convey wrong information. These actions mangle their counterpart's inference process and cause them to adopt false beliefs. We show how different ToM levels adopt different information-theoretic 'attacks'.

4.2. Relevance for animal and machine cognition

This work is relevant for the study of social cognition in artificial (Rabinowitz et al., 2018; Jaques et al., 2019) and biological systems. For example, it adds a model-based reinforcement learning and information theoretic perspective to Goodhart's law (Goodhart, 1984), which states that people tend to try to game statistical regularities used by authorities (e.g., the government or a retailer) for control purposes (e.g., taxation or dynamic pricing). There are many avenues for further enquiry. Future research will have to investigate how closely humans (Ransom et al., 2017; Barnett et al., 2021), or other animals (Premack & Woodruff, 1978), actually follow our theoretic analyses. This is particularly interesting because theory of mind based deception like we discuss here has been proposed to underlie a litany of behaviors, from how the allies planned D-Day (Crawford, 2003) to how birds strategically hide their food cache (Clayton et al., 2007; Emery & Clayton, 2001).

One crucial limitation for the biological plausibility of our simulations is the high computational cost of deep recursive reasoning. This recursion forces a ToM(k) agent to solve $|M|^k$ POMDP problems where M is the number of possible agent models. In Gmytrasiewicz & Doshi (2005) this complexity is evaluated to be PSPACE-hard for finite time horizon. More heuristic approaches like the handcrafting employed by Strouse et al. (2018) may present a more computationally kind solution to this problem. Additionally, model-free rather than planning might give rise to similar dynamics (Dolan & Dayan, 2013). In essence, the Q-values which we here computed via model-based planning may also be approximated via model-free agents that play with agents of lower levels of recursion - although this will substitute time/sample for space complexity.

Theory of mind also shares key overlaps with metacognition, that is an agent's ability to introspect (Schulz et al., 2023; Fleming & Daw, 2017; Carruthers, 2008). Such introspection has been used in artificial intelligence to produce agents that can justify their own behavior (Roy et al., 2022), or large language models that curb their overconfidence (Mielke et al., 2022), and thus has key implications for explainable AI.

4.3. Ethical and AI safety concerns

Our work also offers words of caution for systems with metacognition and theory of mind, particularly as the latter's existence is currently heavily debated with regards to large language models (Sap et al., 2023; Ullman, 2023; Kosinski, 2023). These debates mainly center around whether LLMs possess what we would at most consider ToM(0), making relatively straightforward inferences about the mental state of others but not using them for planning in (semi-)competitive scenarios. LLM behavior has also been studied in more competitive game theoretic settings (Guo, 2023). For example Akata et al. (2023) investigated the behavior of LLMs in repeated games. Their results show how theory of mind like prompts can improve coordination and that GPT-4 has a tendency to play in an unforgiving manner. LLMs have also been coupled with more explicit planning strategies that share similarities with ours, for example for gameplay in Diplomacy (Meta Fundamental AI Research Diplomacy Team et al., 2022). We note however that ”Cicero” is explicitly barred from lying and deceit.

Our work explicitly shows how theory of mind can – without further handcrafting – give rise to deceitful and skeptical agents in a symbolic negotiation task. We note how our setting is a minimal representation of many semi-adversarial human-AI interactions, for example in recommender systems, dynamical pricing or human-LLM conversations. As such, the emergence of theory of mind capabilities will have to be carefully monitored to understand how LLMs reveal what they know and want, and how they interpret what other agents tell them.

A crucial aspect of the safety and ethics perspective on our results is that it is not theory of mind alone that gives rise to deception and skepticism. Rather, what produces this behavior is the coupling of theory of mind with a value function that mis-aligns the utilites of two agents. Our work therefore speaks to the alignment problem in AI safety. For example, the deceitful hiding of intentions may be crucial for the off-switch game (Hadfield-Menell et al., 2017). Given the (model-based) reinforcement learning setting of our task, our work also bears relevance to reinforcement learning from human feedback where both training signals and the learning itself may become skeptical or deceiving.

4.4. Extensions and future work

Our work leaves much opportunity for further investigation. For simplicity, we here limited our paradigm to a two-player setting. Additional research may include further agents, which might result in interesting dynamics as highlighted by Sclar et al. (2022) in a cooperative setting. Providing both the seller and the buyer with richer (item or pricing) options might also make the deception more subtle, and complex.

Conceptually, we also note several straightforward extensions: First, a key factor of this deceptive dynamic is full observability of the actions of the players. Hence, future work might explore the information-seeking perspective of this model (Schulz et al., 2023; Wang et al., 2021), asking how much the buyer is willing to pay, or be paid, to disclose its action and how much the seller is willing to pay to uncover the buyer’s action. Second, certain settings of our task—like the extreme corners of the maze—encourage more or less deceptive behaviours. Future work might thus allow the seller to control the set-up of the maze as part of its planning to maximize utility. We note that this control of the decision environment has links to the game theoretic work on preference elicitation and mechanism design (Becker et al., 1964; Roth, 2008). Third, potential future work adopts a macroeconomic perspective and explores the multi-seller, multi-buyer case. In this setting, sellers need to make inferences about the actions of their competitors as well as about the preferences of their clients. Finally, our agents merely communicate via their (incentivised) actions. More general games of our sort may allow agents to exchange less costly agents, in key overlap with the economic literature on cheap talk (Farrell & Rabin, 1996; Franke et al., 2012).

Acknowledgements

We would like to thank Rahul Bhui and Stefan Bucher for helpful comments. This research has been partly funded by Israel Science Foundation grant #1340/18 (NA; JR), by the Max Planck Society (NA, LS, PD) and the Humboldt Foundation (PD). PD is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764 and of the Else Kröner Medical Scientist Kolleg ”ClinbrAIn: Artificial Intelligence for Clinical Brain Research”.

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. Playing repeated games with Large Language Models, May 2023. URL https://arxiv.org/abs/2305.16867v1.

Alon, N., Schulz, L., Dayan, P., and Rosenschein, J. A (dis-)information theory of revealed and unrevealed preferences. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, November 2022. URL https://openreview.net/forum?id=vcpQW_fGaj5.

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.

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):1–10, 2017. ISSN 23973374. doi: 10.1038/s41562-017-0064. URL http://dx.doi.org/10.1038/s41562-017-0064. Publisher: Macmillan Publishers Limited, part of Springer Nature.

Barnett, S. A., Hawkins, R. D., and Griffiths, T. L. A pragmatic account of the weak evidence effect. *arXiv preprint arXiv:2112.03799*, 2021.

Becker, G. M., Degroot, M. H., and Marschak, J. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964. ISSN 1099-1743. doi: 10.1002/bs.3830090304. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830090304. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830090304.

Berke, M. and Jara-Ettinger, J. Thinking about thinking through inverse reasonin. Technical report, 2021. URL https://doi.org/10.31234/osf.io/r25qn. Type: article.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. A cognitive hierarchy model of games*. 119(3):861–898, 2004. ISSN 0033-5533. doi: 10.1162/0033553041502225. URL https://doi.org/10.1162/0033553041502225.

Camerer, C. F., Ho, T.-H., and Chong, J. K. A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3:157–162, 2015.

Carruthers, P. Meta-cognition in animals: A skeptical look. *Mind & Language*, 23(1):58–89, 2008. Publisher: Wiley Online Library.

Clayton, N. S., Dally, J. M., and Emery, N. J. Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480): 507–522, February 2007. doi: 10.1098/rstb.2006.1992. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2006.1992. Publisher: Royal Society.

Crawford, V. P. Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions. *American Economic Review*, 93(1):133–149, March 2003. ISSN 0002-8282. doi: 10.1257/000282803321455197. URL https://www.aeaweb.org/articles?id=10.1257/000282803321455197.

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., and Camerer, C. In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6):1222–1231, 2013.

Dolan, R. J. and Dayan, P. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013. ISSN 08966273. doi: 10.1016/j.neuron.2013.09.007. URL http://dx.doi.org/10.1016/j.neuron.2013.09.007. Publisher: The Authors ISBN: 1097-4199 (Electronic)\r0896-6273 (Linking).

Doshi, P., Gmytrasiewicz, P., and Durfee, E. Recursively modeling other agents for decision making: A research perspective. 279:103202, 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2019.103202. URL https://www.sciencedirect.com/science/article/pii/S000437021930027X.

Emery, N. J. and Clayton, N. S. Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, 414(6862):443–446, November 2001. ISSN 1476-4687. doi: 10.1038/35106560. URL https://www.nature.com/articles/35106560. Number: 6862 Publisher: Nature Publishing Group.

Farrell, J. and Rabin, M. Cheap Talk. *Journal of Economic Perspectives*, 10(3):103–118, September 1996. ISSN 0895-3309. doi: 10.1257/jep.10.3.103. URL https://www.aeaweb.org/articles?id=10.1257/jep.10.3.103.

Fleming, S. M. and Daw, N. D. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91–114, 2017. ISSN 0033295X. doi: 10.1037/rev0000045.

Franke, M., De Jager, T., and Van Rooij, R. Relevance in Cooperation and Conflict*. *Journal of Logic and Computation*, 22(1):23–54, February 2012. ISSN 0955-792X. doi: 10.1093/logcom/exp070. URL https://doi.org/10.1093/logcom/exp070.

Gmytrasiewicz, P. J. and Doshi, P. Interactive POMDPs: Properties and preliminary results. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004*, 3(July 2004):1374–1375, 2004. doi: 10.1109/AAMAS.2004.154. ISBN: 1581138644.

Gmytrasiewicz, P. J. and Doshi, P. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005. ISSN 1076-9757. doi: 10.1613/jair.1579. URL https://jair.org/index.php/jair/article/view/10414.

Goodhart, C. A. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pp. 91–121. Springer, 1984.

Guo, F. GPT Agents in Game Theory Experiments, May 2023. URL http://arxiv.org/abs/2305.05516. arXiv:2305.05516 [econ, q-fin].

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. The Off-Switch Game, June 2017. URL http://arxiv.org/abs/1611.08219. arXiv:1611.08219 [cs].

Ho, M. K., Saxe, R., and Cushman, F. Planning with Theory of Mind. *Trends in Cognitive Sciences*, 26(11):959–971, November 2022. ISSN 1879307X. doi: 10.1016/j.tics.2022.08.003. Publisher: Elsevier Ltd.

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. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004254. URL https://dx.plos.org/10.1371/journal.pcbi.1004254.

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, October 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2019.04.010. URL https://www.sciencedirect.com/science/article/pii/S2352154618302055.

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. ISSN 13646613. doi: 10.1016/j.tics.2016.05.011. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661316300535.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kopp, C., Korb, K. B., and Mills, B. I. Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to ”fake news”. *PLOS ONE*, 13(11):e0207383, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207383. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207383. Publisher: Public Library of Science.

Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models, March 2023. URL http://arxiv.org/abs/2302.02083. arXiv:2302.02083 [cs].

Meta Fundamental AI Research Diplomacy Team, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, August 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00494. URL https://doi.org/10.1162/tacl_a_00494.

Miura, S. and Zilberstein, S. A unifying framework for observer-aware planning and its complexity. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 610–620. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/miura21a.html.

Ng, A. Y. and Russell, S. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*, pp. 663–670. Morgan Kaufmann, 2000.

Oey, L. A., Schachner, A., and Vul, E. Designing and Detecting Lies by Reasoning About Other Agents. *Journal of Experimental Psychology: General*, 2022. ISSN 00963445. doi: 10.1037/xge0001277. Publisher: American Psychological Association.

Oguntola, I., Hughes, D., and Sycara, K. Deep interpretable models of theory of mind for human-agent teaming. 2021. URL http://arxiv.org/abs/2104.02938.

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

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.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 2586–2591, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Ransom, K., Voorspoels, W., Perfors, A., and Navarro, D. A cognitive analysis of deception without lying. Cognitive Science Society, 2017.

Ray, D., King-Casas, B., Montague, P., and Dayan, P. Bayesian model of behaviour in economic games. *Advances in neural information processing systems*, 21, 2008.

Roth, A. E. Deferred acceptance algorithms: history, theory, practice, and open questions. *International Journal of Game Theory*, 36(3):537–569, March 2008. ISSN 1432-1270. doi: 10.1007/s00182-008-0117-6. URL https://doi.org/10.1007/s00182-008-0117-6.

Roy, N. A., Kim, J., and Rabinowitz, N. Explainability Via Causal Self-Talk. *Advances in Neural Information Processing Systems*, 35:7655–7670, December 2022.

Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., and Gläscher, J. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488, 2020. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2020.107488. URL https://www.sciencedirect.com/science/article/pii/S0028393220301597.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, April 2023. URL http://arxiv.org/abs/2210.13312. arXiv:2210.13312 [cs].

Schulz, L., Fleming, S. M., and Dayan, P. Metacognitive Computations for Information Search: Confidence in Control. *Psychological Review*, 2023. doi: 10.1037/rev0000401. URL https://doi.org/10.1037/rev0000401.

Sclar, M., Neubig, G., and Bisk, Y. Symmetric machine theory of mind. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19450–19466. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sclar22a.html.

Silver, D., Singh, S., Precup, D., and Sutton, R. S. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103535. URL https://www.sciencedirect.com/science/article/pii/S0004370221000862.

Strouse, D., Kleiman-Weiner, M., Tenenbaum, J., Botvinick, M., and Schwab, D. J. Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Tejwani, R., Kuo, Y.-L., Shu, T., Katz, B., and Barbu, A. Social interactions as recursive mdps. In *Conference on Robot Learning*, pp. 949–958. PMLR, 2022a.

Tejwani, R., Kuo, Y.-L., Shu, T., Stankovits, B., Gutfreund, D., Tenenbaum, J. B., Katz, B., and Barbu, A. Incorporating rich social interactions into mdps. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7395–7401. IEEE, 2022b.

Ullman, T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023. URL http://arxiv.org/abs/2302.08399. arXiv:2302.08399 [cs].

Wang, Y., Zhong, F., Xu, J., and Wang, Y. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. *arXiv preprint arXiv:2111.09189*, 2021.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., and Montague, P. R. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, 8(12):e1002841, 2012.

Xiang, Y., Vélez, N., and Gershman, S. J. Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, pp. No Pagination Specified–No Pagination Specified, 2023. ISSN 1939-2222. doi: 10.1037/xge0001336. Place: US Publisher: American Psychological Association.

Zaslavsky, N., Hu, J., and Levy, R. P. A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*, 2020.

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. Modeling consumer learning from online product reviews. *Marketing Science*, 32(1):153–169, 2013. ISSN 1526548X. doi: 10.1287/mksc.1120.0755.

A. Appendix.

A.1. Optimal pricing by ToM(0) seller

The below outlines the full computations involved in computing the optimal price by the ToM(0) seller.

$$\begin{aligned} m^*_{k=0}(i_3, a_1) &= \underset{m(i)}{\text{argmax}} \{\, E[U_S(i)|a_1]\} \\ &= \underset{m(i)}{\text{argmax}} \left\{ \int_R m(i) \cdot \hat{P}_{k=-1}(a_3 = i|r(i), m(i)) p_{k=0}(r(i)|a_1 = i_1) dr(i) \right\} \end{aligned} \tag{12}$$

A.2. Q-value computation of ToM(1) buyer

Here, we present the ToM(1) buyer's Q-value computation in full:

$$\begin{aligned} Q^{t=1}_{k=1}(a_1 = i_1, \mathbf{d}) &= U^{t=1}_B(i_1, \mathbf{d}) + E[U^{t=3}_B(i_3, m^*_{k=0}(i_3, a_1))|a_1 = i_1] \\ &= U^{t=1}_B(i_1, \mathbf{d}) + \sum_{i_3} U^{t=3}_B(i_3, m^*_{k=0}(i_3, a_1)) \hat{P}(a_3 = i_3|m^*_{k=0}(i_3, a_1), a_1 = i_1) \end{aligned} \tag{13}$$