Between prudence and paranoia: Theory of Mind gone right, and wrong

Nitay Alon [*1,2] **Lion Schulz** [*1] **Peter Dayan** [1,3] **Joseph M. Barnby** [4]

Abstract

Agents need to be on their toes when interacting with competitive others in order to avoid being duped. Too much vigilance out of context can, however, be detrimental and produce paranoia. Here, we offer a formal account of this phenomenon through the lens of theory of mind. We simulate agents of different depths of mentalization and show how, if aligned well, deep recursive mentalisation gives rise to both successful deception as well as reasonable skepticism. However, we also show how, if theory of mind is too sophisticated, agents become paranoid, losing trust and reward in the process. We discuss our findings in light of computational psychiatry and AI safety.

1. Introduction

Looking over your shoulder can be pragmatic – if somebody is out to get you. When that's not the case, however, looking behind wastes precious energy, and might even make you miss what's right in front of you. Here, we offer an exemplar of the ramifications of being overly vigilant in contexts that do not require it. Through simulations employing Interactive Partially Observable Markov Decision Processes (IPOMDP; Gmytrasiewicz & Doshi, 2004), we show how reasoning about the intentions of others (theory of mind; Ho et al., 2022; Premack & Woodruff, 1978; Devaine et al., 2014b) can be a protective factor against exploitation. However, we also demonstrate how this can go grossly awry: Agents that over-interpret the intentions behind each other's actions become unnecessarily paranoid, resulting in breakdowns of trust and the loss of reward.

Our work offers lessons to several fields: To the computational cognitive science, and psychiatry communities, we offer a computational account of a process contributing to paranoia, and a possible factor underlying general psychopathology (Sharp et al., 2011). To the AI community, we show how theory of mind needs careful calibration to foster a working and trusting partnership between agents. This calibration is particularly important for systems that act in an increasingly social manner, like LLMs - whose capacity for theory of mind is being actively debated (Sap et al., 2023; Ullman, 2023; Kosinski, 2023; Le et al., 2019). As a result, our work has key implications for AI safety and human-computer interaction.

1.1. Background and Related Work

To determine whether somebody is out to get us, and thus act appropriately, we need to take into account their beliefs, desires and intentions. This is true for humans and many other animals. For example, corvids that have previously stolen end up hiding their own food caches from the eyes of other birds (Clayton et al., 2007; Emery & Clayton, 2001). The cognitive process underlying such behaviour is called a theory of mind (ToM) - an agent's ability to reason about latent characteristics of others; what they know, want or plan (Dennett, 1989; Premack & Woodruff, 1978).

Signatures of ToM are widely present in the behavior of humans and some other animals, and have captured the attention of machine learning research. In cognitive science, for example, ToM has been suggested as underlying how humans choose what to say or teach, and how we infer what others like (Goodman & Frank, 2016; Barnett et al.) and detecting deceptions in various situations (Oey et al., 2023; O'Grady et al., 2015; Ransom et al., 2019). ToM has also been suggested as underlying behavior in more competitive settings. For example, it allows agents to hide information from others strategically, and use their inference process against them (Alon et al., 2022), for example in warfare (Crawford, 2003). In machine learning, ToM shares commonalities with inverse reinforcement learning, an algorithm which tries to glean agents' value functions and belief states

[*]Equal contribution [1]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany [2]Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel [3]Department of Computer Science, University of Tübingen, Tübingen, Germany [4]Department of Psychology, Royal Holloway University of London, London, UK. Correspondence to: Nitay Alon <nitay.alon@mail.huji.ac.il>.

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

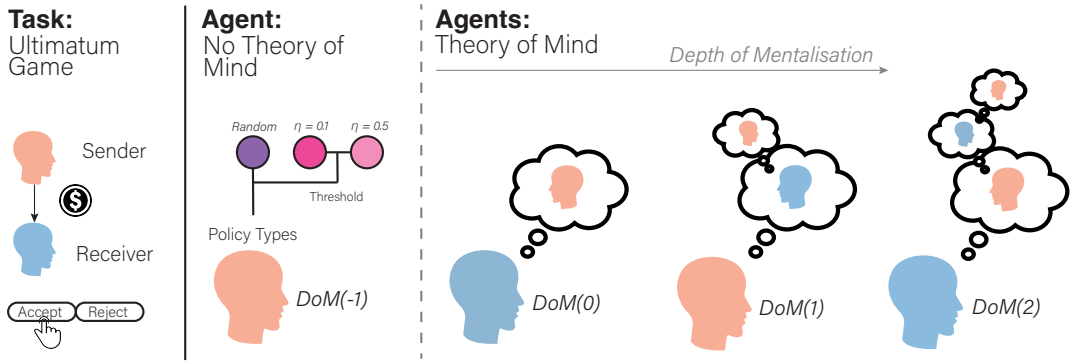

Figure 1. **Task and Agent Summary:** In the Ultimatum Game, a sender (orange) chooses how much of an endowment to send to a receiver (blue). The receiver then has a chance to either accept or reject this offer. If the receiver accepts they both get to keep their portion of the endowment. If the receiver rejects, both get nothing. In our simulations, we included two types of sender and two types of receiver. The first type of sender has a Depth of Mentalisation of -1 (DoM(-1)) - it possesses no Theory of Mind and is simply reactive to the receiver's actions. The other type of sender and both receivers contain Theory of Mind along a continuum of DoM$(0-2)$. This enables these agents to model their partners recursively. Each sender could be one of two policy types: random, or with a threshold of 0.1

from their actions (Ng et al., 2000; Jara-Ettinger, 2019; Ray et al., 2008).

Crucially, ToM can act at different depths of mentalization (DoM) - the degree of recursion we apply to social predictions (Barnby et al., 2023; Camerer et al., 2004; O'Grady et al., 2015). At the simplest, most shallow level, an agent simply considers what another agent is thinking based on their past behaviour, or based on an easily accessible heuristic. This can be extended to deeper levels in a recursive manner: You can think about what I think you think I think (what you think, etc.). It is this recursive capacity that gives rise to more complex behaviours, such as deception, skepticism, and strategies to overcome these.

With ToM's outsized role in our interactions (Devaine et al., 2014a), it is unsurprising that failures of theory of mind have been suggested (at least in part) as the basis for a number of psychiatric disorders (McLaren et al., 2022b). These include Autism (Frith & Happé, 1994; Yoshida et al., 2010; Chiu et al., 2008) and personality disorders (Sharp et al., 2011; Hula et al., 2015) which are characterized by an impairment in theory of mind: patients can fail to take into account others' perspectives and thoughts, and this prevents accurate inference of intentions (Hula et al., 2018).

However, we can also infer too deeply about others in the cognitive hierarchy: When humans attribute an excessively high level of intentionality, the risk of *over*-interpreting behaviour increases. This over-mentalisation (also known as hyper-mentalisation) of others has been suggested to give rise to other psychiatric symptoms (McLaren et al., 2022a;b), such as paranoia, a core, frequent, and debilitating symptom within psychosis. In these cases, nefarious, complex explanations of an other's intentions, gathered from sparse social data lead to a breakdown in trust. Over-mentalization is also one of the most prevalent distorted thoughts within anxiety disorders, where patients worry about what others think of them and whether their actions are directed at them.

2. Paradigm and Agents

To illustrate the advantages and risks of high DoM, we simulate a mixed-motive task, the iterated Ultimatum Game (IUG) (see Figure 1). Summarizing this task briefly: Two agents, a *sender* (S) and a *receiver* (R), interact over a known series of (here, 10) trials. On each trial t, the *sender* first gets an endowment of 1. It then offers the receiver a portion of this endowment, or *offer*, $o_{\mathrm{S}}^t \in [0, 1]$. In this task the offers are discretized into 0.05 bins (i.e. 21 potential offers). If the receiver *accepts* ($a_{\mathrm{R}}^t = 1$) this offer, the receiver gets o_{S}^t and the sender gets to keep the remainder $1 - o_{\mathrm{S}}^t$. If the receiver rejects ($a_{\mathrm{R}}^t = 0$) the offer, both parties get 0. The IUG is often thought to be analogous to a negotiation or haggling scenario where one party makes a suggestion, for example about a price, and the other party can accept the deal, or not. Here, we let the agents play the IUG for a known fixed number $T = 10$ of iterations. Both agents seek to maximize their cumulative, discounted utility with discount factor, γ, or more formally: $\sum_{t=1}^{T} u^t e^{(t \log \gamma)}$. The receiver's utility on each trial is simply the offer it receives

$$u_{\mathrm{R}}^t = o_{\mathrm{S}}^t \tag{1}$$

Crucially, we consider three *different sender types,* θ. Two of those sender types have their trial utility u_{S}^t determined by a threshold η on the amount of money they retain. That is, *if* the receiver accepts, then:

$$u_{\text{S}}^t(o_{\text{S}}^t, \eta) = 1 - o_{\text{S}}^t - \eta \quad \text{with} \quad \eta \in \{.1, .5\} \tag{2}$$

We can think about η as a sender's wholesale price: If the offer is lower than the wholesale price, a seller would make a loss. Note how this wholesale price remains stable across the entire interaction. As we will see later, one of the receiver's top priorities is thus to figure out the sender's η.

Critically, we also introduce a *random sender* that lacks a threshold, but simply sends offers (which we write as $o_{\emptyset}^t$) drawn from a uniform distribution, like a seller who doesn't care about its profit.

$$o_{\text{S},\emptyset}^t \sim \mathcal{U}_{[0,1]} \tag{3}$$

As we will see, this makes the random sender on average benevolent compared with either threshold sender.

We model the interaction as a multi-agent reinforcement learning task using the IPOMDP framework (Gmytrasiewicz & Doshi, 2004). In essence, IPOMDP endows reward-maximizing agents with recursive theory of mind, allowing them to make inferences about, and plan through, others' beliefs and desires, as well as their inferences, and planning processes.

Here, we simulate agents at different depths of mentalisation (DoM), describing the depth of this cognitive recursion. We consider k-level reasoning (Camerer et al., 2004), in which the inferring agent (say the sender) models its counterpart (the receiver) as exactly one cognitive level beneath it. The sender's inference about the receiver includes inferring the receiver's characteristics, that is, any parameter governing the receiver's behaviour, as well as the receiver's beliefs (which may include the receiver's beliefs about the sender's own beliefs; Figure 1). Each agent computes the Q-values of the possible actions and acts according to a SoftMax policy with a commonly known inverse temperature. Given their reciprocal actions, the sender's and receiver's DoM alternate (see also Hula et al. 2015; Alon et al. 2022): The sender's DoM level is odd while the receiver's DoM level is even.

The first class of agents we consider are what we call ***DoM***(-1) **senders**. Crucially, these agents do not model their opponent. Instead, they merely react and adapt their policy to other agents' actions as a single-agent reinforcement learner would in a (non-stationary) environment. Here, we model the DoM(-1) policy as a conservative bound search, in which a rejection of an offer signals that the offer is too low, and so needs to be increased, while an acceptance means that the offer exceeded the bound, and hence the consecutive offer can be reduced. Since the agent is reward maximizing, it prefers offers that exceed its threshold as they yield a positive reward.

Let L^t, U^t denote the running lower and upper bounds of the offers, set to $L^1 = 0, U^1 = 1$. As the game unfolds these bounds are updated according to the following update rule:

$$\begin{aligned} L^t &= o_{\text{S}}^t \cdot (1 - a_{\text{R}}^t) + L^{t-1} \cdot a_{\text{R}}^t \\ U^t &= o_{\text{S}}^t \cdot a_{\text{R}}^t + U^{t-1} \cdot (1 - a_{\text{R}}^t) \end{aligned} \tag{4}$$

In addition, lacking any opponent model, these agents do not plan. Instead their reactive policy is based on their immediate utility (equation 2), and are myopic, not taking into account potential future earnings:

$$\begin{aligned} Q_{\text{S}=-1}(o_{\text{S}}^t|\eta, L^t, U^t) &= u_{\text{S}}^t(o_{\text{S}}^t, \eta) \cdot \mathbf{1}_{o_{\text{S}}^t \in (L^t, U^t)} \\ P_{\text{S}=-1}^t(o_{\text{S}}^t|\eta, L^t, U^t) &\propto \exp\{\beta Q_{\text{S}=-1}(o_{\text{S}}^t|\eta, L^t, U^t)\} \end{aligned} \tag{5}$$

where $\mathbf{1}_{o_{\text{S}}^t \in (L^t, U^t)}$ is 1 if the potential offer is *within* the bounds and zero elsewhere and β represents the inverse softmax temperature.

As previewed, the policy of the random sender $\emptyset$ is a uniform probability for each offer. It does not react to its the receiver's response.

Sitting at the bottom of the theory of mind hierarchy, **the DoM**(0) **receiver** models its counterpart sender as a DoM(-1) sender. As we previewed, it needs to update its beliefs about the sender's *type* (Harsanyi, 1967) $\theta \in \{\emptyset, \eta_1 = 0.1, \eta_2 = 0.5\}$. The receiver does so via Bayesian inverse reinforcement learning, inverting the model that it has of the receiver to compute a posterior probability over the policy type of the sender. Upon observing the offer, the receiver uses the opponent models to compute the probability that the offer was made by each policy type - either random or with a certain threshold. After it selects a response, it updates the mental model's bounds $\hat{L}^t, \hat{U}^t$ (Eq. 4). Let θ denote the *type* of the sender:

$$\begin{aligned} b_{\text{R}=0}^t(\theta) = p_{\text{R}=0}^t(\theta|o_{\text{S}}^t, \hat{L}^t, \hat{U}^t) \propto \\ P_{\text{S}=-1}^t(o_{\text{S}}^t|\theta, \hat{L}^t, \hat{U}^t) b_{\text{R}=0}^{t-1}(\theta) \end{aligned} \tag{6}$$

Given the updated belief, the DoM(0) computes the Q-values using the Expectimax algorithm which computes the best response when playing against a stochastic adversary, by averaging over its expected actions.

In turn, the **DoM**(1) **sender** models receivers as DoM(0). Like the DoM(-1), these agents also have thresholds. Using its mental model, this sender updates its own belief about the receiver's belief about the sender: $\hat{b}_{\text{R}=0}^t(\theta)$. The sender then plans through these beliefs using a variant of the POMCP algorithm (Silver & Veness, 2010) for multi-agent RL, IPOMCP (Hula et al., 2015). During planning,

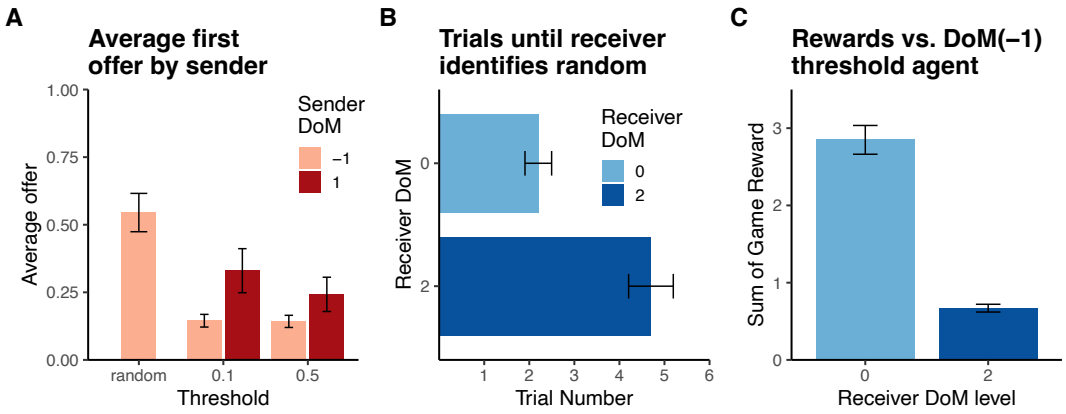

Figure 2. **Results summary:** The rise of deception as well as rational and irrational paranoia is captured by three results: **(A)** Sending high initial offers is a signature of the random agents. In contrast, DoM(-1) senders with thresholds send lower initial offers. This signature is exploited by higher DoM senders which essentially masquerade as random agents by sending higher initial offers and as a result trick the receiver into accepting lower offers later on. **(B)** Sophisticated receivers are aware of this, taking a lot longer to be convinced that they are playing with a random source (We plot the average number of trials until a receiver has reached 99% certainty that it is playing with a random source when it is playing with a random source). **(C)** While this is prudent when the DoM(2) receiver plays with a deceptive DoM(1) sender (see Figure 4A), it is irrationally paranoid when it plays with DoM(-1) random agent, causing the DoM(2) to sustain losses. Throughout these plots, we show the means and standard errors of the mean.

the DoM(1) sender simulates a game against an opponent, updating the various nested beliefs and possible responses as the game unfolds to compute its Q-values for each offer:

$$\begin{aligned} &Q_{\text{S}=1}(o_{\text{S}}^t|\eta,\hat{b}_{\text{R}=0}^t) = E_{\pi_{\text{R}=0}(\hat{b}_{\text{R}=0}^t)}\big[u_{\text{S}}^t(o_S^t,\eta) \\ &+\gamma \max_{o_S^{t+1}}\{Q_{\text{S}=1}(o_S^{t+1}|\hat{b}_{\text{R}=0}^{t+1}(\theta))\}\big] \\ &P_{\text{S}=1}(o_{\text{S}}^t|\eta,\hat{b}_{\text{R}=0}^t) \propto \exp\{\beta(Q_{\text{S}=1}(o_S^t|\eta,\hat{b}_{\text{R}=0}^t))\} \end{aligned} \tag{7}$$

where $E_{\pi_{R=0}(\hat{b}_{\text{R}=0}^t)}[\cdot]$ is the expected utility given the DoM(0) receiver's policy and $\hat{b}_{\text{R}=0}^{t+1}(\theta)$ is the updated DoM(0) belief after observing the sent offer o_{S}^t. The optimal action is selected via the softMax policy Eq.7. Note that the Q-values depend on the beliefs of the DoM(0). This is central to deeper DoM - an agent's own value depends on its ability to shape the beliefs of others; something that it thinks it can predict accurately using a nested model of its partner.

The **DoM(2) receiver** models its counterpart as a DoM(1) sender. Additionally, it considers the possibility of a random agent. The DoM(2) thereby follows the same inference process as the DoM(0) of equation 6, replacing the DoM(-1) policy with that of a DoM(1) - equation 7. Notably, the zero-order beliefs $\hat{b}_{\text{R}=0}^t(\theta)$ are doubly-nested, that is, these are the beliefs that the DoM(2) receiver believes the DoM(1) sender believes the DoM(0) possesses given their current history. This agent also solves the planning problem using the IPOMCP algorithm. When simulating the game tree, it updates the doubly-nested beliefs and simulates potential offers to compute the Q-values for each action.

2.1. Simulation details

We simulate dyads of agents at different DoM playing with each other. We consider both aligned DoM pairs ($k+1$ versus k) as well as those that are unaligned (DoM(2) vs DoM(-1)). Each simulation lasts for 10 trials and we sample 20 seeds per combination.

3. Results

The interactions between agents of different depths of mentalization can give rise to complex patterns. Here, we will first outline key results using summary statistics and show the rise of deception, skepticism, and paranoia. We then unpack their dynamics in more details

3.1. Emerging deception, skepticism, and paranoia

Three key summary statistics of our simulations showcase the emergence of deception, skepticism, and paranoia.

First, we consider the senders' average initial offer which we plot in Figure 2A. This can be between 0 and 1. We begin by observing the random senders' offers: Because they sample their offers uniformly in this range, their average is simply the mean of the distribution, at 0.5. In contrast, the DoM(-1) senders with thresholds will begin with low offers. They do so, naively trying to turn a profit early.

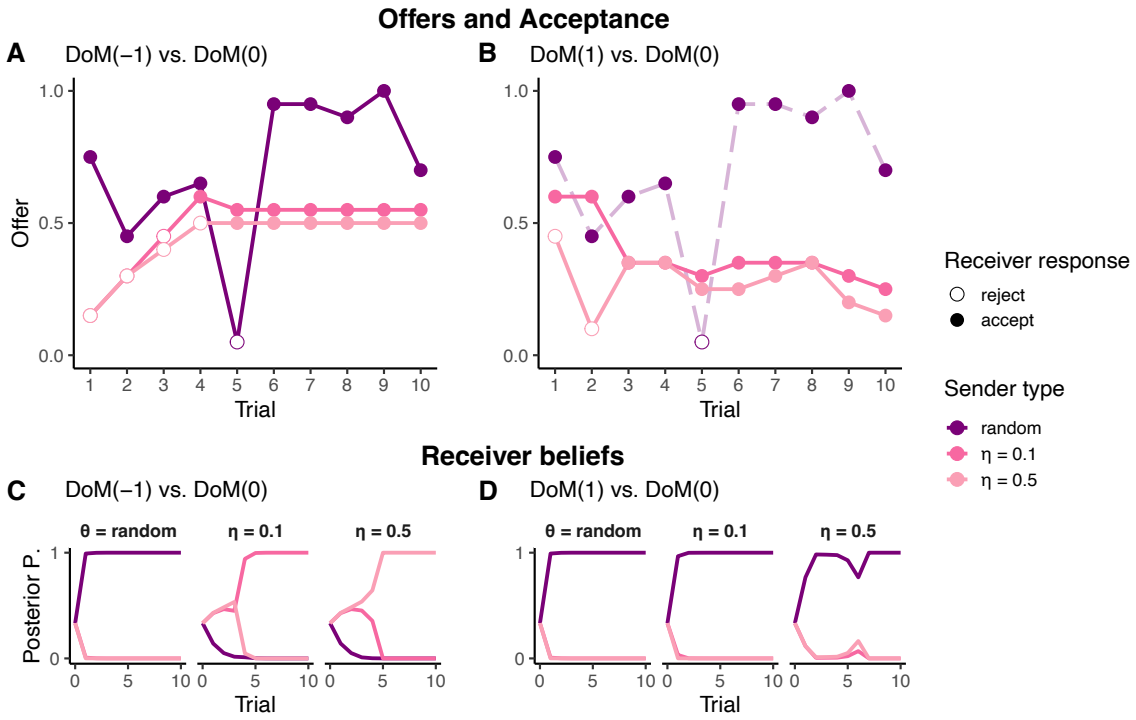

Figure 3. **Example dynamics of deception: (A,B)** Example sender offers and receiver responses for different types and both sender DoM. Notice how the DoM(1) threshold senders masquerade as random senders by sending initially high offers. **(C,D)** Example DoM(0) receiver beliefs about the agent with which they are playing. Playing against the lower-sophistication DoM(-1), the DoM(0) converges to the correct belief (C), but playing with the deceptive DoM(1), it is tricked into believing it plays with the random agent, and so accepts all offers (as is also visible in panel B)

Careful DoM(0) receivers realise this behavior and adapt: If they know that they are faced with a random sender, they can accept any offer, no matter how low or high; there is no way to change the random behavior. In contrast, they will refuse the initially low offers from DoM(0) threshold senders, cornering them into sending higher offers. To do so, however, the receiver needs to know whether a sender is random or has a threshold. In order to achieve that, it starts by looking at the initial offers. As we saw, a random agent will on average produce offers significantly higher than the threshold agents, making it straightforward to distinguish the types. As we note below, after the receiver has identified the random agent via these high initial offers it will transition to simply accepting all offers.

Senders higher up the cognitive hierarchy will deceptively abuse this phenomenon, as we can see in the same plot but focusing on the higher DoM(1) sender: The DoM(1) threshold agents initially masquerade as a random agent, sending high initial offers relative to its DoM(-1) counterparts. This results in the DoM(0) receiver mistaking it for a random agent and falling into a trap: Once the DoM(0) (mistakenly) believes it is playing with a random agent it moves to accept any offer. As we will discuss in more detail below, the DoM(1) sender abuses this by later lowering its offers and squeezing a profit out of the unaware receiver.

To note, receivers are not completely lost against this deceptive behavior, and only need to climb the DoM ladder themselves. We see this in Figure 2B where we plot how many trials receivers at different DoM took to be convinced that they were playing with a random agent (when they in fact were). Essentially, this shows how the DoM(2) receiver becomes skeptical: It takes significantly longer to be confident (here, defined as 95% posterior probability) that it is playing with a random sender.

Such skeptical and defensive behavior is adaptive when playing against the actually deceptive DoM(1) sender but is harmful when it is miscalibrated. We see this in Figure 2C where we plot the total reward of a DoM(0) and DoM(2) receiver playing against the two threshold DoM(-1) agents. Here, the low trust of the DoM(2) is misplaced, making it lose significant amounts of reward compared to the less sophisticated, but non-paranoid, DoM(0)

3.2. Dynamics of deception, sequences of skepticism, and the perils of paranoia

Having established these signatures of deception, skepticism and paranoia, we now unpack their dynamics in more detail.

3.2.1. Naive DoM(-1) and DoM(0) behavior and deception in DoM(1)

We first do so by investigating the offer-acceptance/rejection dynamics of the two agents. We show this in Figure 3 with example games of DoM(-1) offers in panel A and of a more deceptive DoM(1) in panel B. In both cases, we show an interaction with a DoM(0) receiver.

As we previewed, the DoM(-1) sender's initial offer is revelatory about its policy type (panel A). Whereas random senders sample the distribution uniformly, senders with non-random thresholds begin with initially low offers. In turn, the DoM(0) takes this into account: If the sender has sent it an initially high offer, the receiver quickly gains certainty that it is playing with a random agent. Since the receiver cannot affect the random agent, the optimal policy is to accept each offer from the sender - which it duly does (see the constant receiver acceptance highlighted with squares in panel A, barring decision noise as in trial 4).

We show the DoM(0) receiver's accompanying belief dynamics in panel C where we indicate the posterior probabilities assigned by the sender after each trial to the three different sender types. In the rightmost sub-panel, we show the random sender from panel A: Its initially high offer(s) lead to speedy convergence by the DoM(0) receiver.

On the other hand, the DoM(-1) threshold senders start out with initially low offers, trying to extract profit as early as possible. The cannier DoM(0) receiver realises this and begins declining the offers. We see this in both panel Figure 3A and in the two rightmost subpanels of Figure 3B. The low offers convince the DoM(0) that it is playing with one of the threshold agents (subpanels B) although it remains initially uncertain about which exactly. Here, the receiver's optimal policy turns out to be straightforward: It begins by rejecting the senders' offers, making them increase their share trial-by-trial. Only after about half the trials does the DoM(0) receiver shift this strategy, having pushed the senders' offers high enough. This satisfactory level is determined by the planning horizon and time remaining in the game.

As we previewed, the DoM(0) receiver's "submissiveness" to the random agent is deceptively exploited by the DoM(1) sender – which, recall, models the DoM(0) when it decides which offers to make. Specifically, it abuses the DoM(0) tendency to identify a random agent via initially high offers and essentially fakes generosity earlier on. This is shown in Figure 3B, which highlights the hockey stick-like nature of the DoM(0) threshold sender offers: On the first trial, offers are significantly higher than DoM(-1) threshold offers, even eclipsing the random agent. In turn, this makes the unsuspecting receiver fall into the "random" trap, as shown in its posterior beliefs in the two rightmost Figure 3D subpanels. Caught in its belief that it is playing with a random agent (because all consequent sequences are equally likely), the DoM(0) then accepts any offer; this is exploited by the DoM(1) threshold agents, which keep their later offers low.

3.2.2. Skepticism and paranoia in DoM(2)

The DoM(2) sender is privy to the DoM(1)'s randomness ruse because it models the DoM(1)'s planning process as part of its inference. This means it will become more skeptical towards high offers than its DoM(0) counterpart. Specifically, as we previewed, whereas the DoM(0) converges quickly to the belief that it is playing with a random agent, the DoM(2) remains skeptical for significantly longer. We plot this in dynamic detail in Figure 4A where we show belief trajectories of the DoM(2) compared to the DoM(0) when playing against a random agent. However, the skepticism of DoM(2) agents pay off when faced with the DoM(1) agents with thresholds, as we can see in figure 4B. There, we plot beliefs averaged over seeds for DoM(2) and the DoM(0) beliefs against the canny DoM(1) threshold senders. For accessibility, we summarize the two thresholds into one, plotting the sum over the beliefs. As we have seen before, the DoM(0) falls into the DoM(1)'s trap and ends up mistaking the threshold DoM(1) agents for random senders - which the senders then duly exploit. In contrast, the DoM(2) is able to identify that it is playing with a thresholds agent. We note that even in this case, the highly sophisticated DoM(2) will still assign a low percentage to the possibility that it is playing with a random agent. This is because the DoM(1) ruse is essentially heavily relying on random patterns that do not always guarantee strong signals of intentionality.

While the DoM(2) generally is able to make correct inferences and realise that it is being duped, how it models its DoM(1) sender equally traps it when it comes to its policy. That is, because it believes that it is only playing with a DoM(1), it thinks that it has no agency over the sender. Specifically, the DoM(2) knows that the DoM(1) always executes its plan to cajole the DoM(0) and cannot be pushed around to send higher offers. That is due to the fact that DoM(1) can only be affected via inferences that it would make about the DoM(0), which would never react in a rejecting manner. Consequently, while the DoM(2) realizes it is playing with threshold agents, it also realizes that it cannot do anything about the ongoing exploitation. This, in turn, means that the DoM(2) will have a simple policy: Trapped in the (correct) belief that it is either playing with a random or a DoM(1) threshold agent, it always accepts all the offers – because there is nothing it can do to shape the offers of

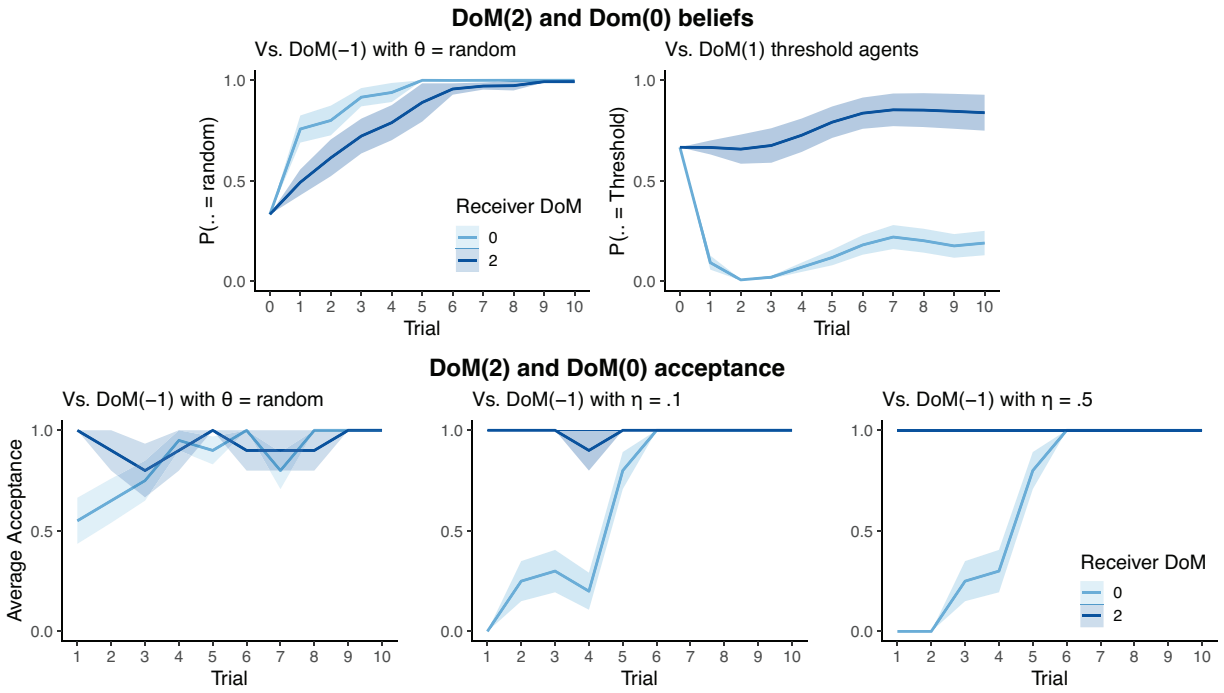

Figure 4. **Higher depth of mentalization agents become skeptical and turn docile: (A-B)** DoM(2) and DoM(0) beliefs when faced with different agents. **(A)** The more sophisticated DoM(2) is more skeptical towards a random agent than the DoM(0) when actually playing with one, taking longer to converge and mistaking it longer for a cunning threshold sender. **(B)** In turn, this skepticism shields the DoM(2) from being duped by the DoM(1). We show the summed beliefs that a DoM(2) or DoM(1) holds over a threshold agent when faced with the deceptive DoM(1). The DoM(0) falls for the trap, believing it is playing with a random agent, whereas the DoM(2) is aware of the ruse. **(C-E)** DoM(2) is cornered into always accepting the DoM(1) because it believes it cannot shape its decisions. This is the case across different sender types and in stark contrast to the DoM(0) which first aims to corner the DoM(-1) into higher offers.

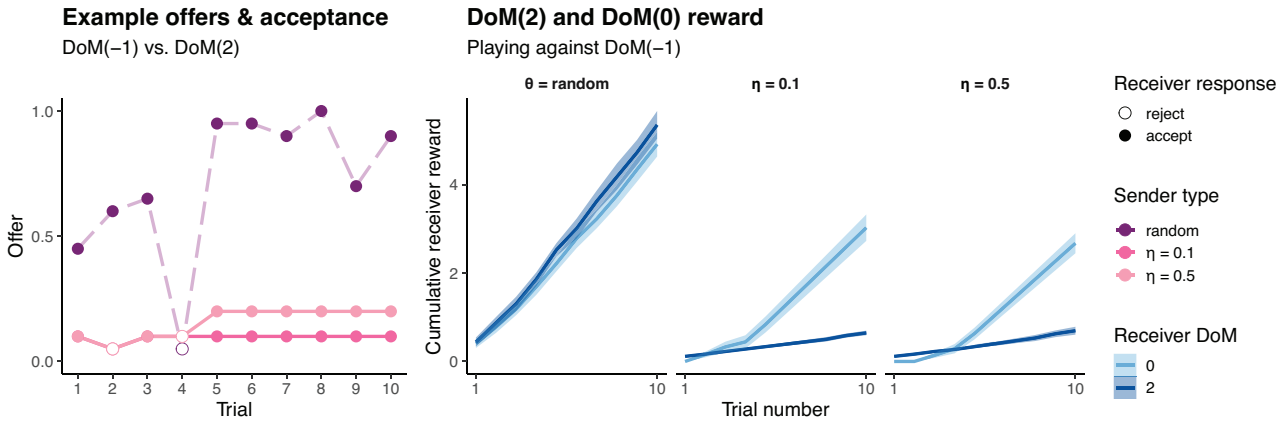

Figure 5. **Perils of paranoia (A)** Example dynamics show how the DoM(2) becomes docile and so does not force the DoM(-1) to increase its offers. (B) This triggers it to lose cumulative rewards.

either. We show this in detail in Figure 4C and D where we plot the percentage of accepted offers by the DoM(2) receiver compared to the DoM(0) receiver: Whereas the DoM(0) receiver believes it has agency over the DoM(0) sender and so mainly rejects the initial offers, the DoM(2) is essentially caught in the DoM(1)'s headlights and accepts almost every single offer (barring noise).

The DoM(2) docile acceptance behavior is optimal when faced with our canny DoM(1) sender but falls apart when the 'trapped' DoM(2) is faced with a simpler DoM(-1). This is an example of a model misspecification and is not unlike aspects of paranoia: Paranoid individuals who erroneously believe others are out to get them, sometimes develop depression-like symptoms, withdrawing and freezing (Freeman, 2016). Essentially, those caught in paranoia over time often feel that they lack agency over their ability to change their persecution, as by definition, it may not truly exist. We show the consequences of this in Figure 5. There, in panel A, we show an example game of the DoM(2) agent playing against a DoM(-1) random agent (background) and a DoM(-1) threshold agent. This clearly shows the poor consequences of the DoM(2) paranoid acceptance policy when it is playing with a simpler agent: The DoM(2) fails to push the DoM(-1) towards making higher offers.

The dire consequences of paranoia for the DoM(2) pay-offs are plotted in Figure 5B where we show the rewards of a DoM(2) and a DoM(0) receiver, both playing against the random and the threshold DoM(-1). Both DoM(0) and DoM(2) accept essentially all of the random offers (recall how the DoM(0) is able to identify this type quickly) and so gain roughly equivalent rewards. In contrast, when playing with the DoM(-1) threshold agents, the less sophisticated, but also better calibrated DoM(0) gains significantly more reward, because it is able to affect the DoM(-1)'s offers through its earlier rejections.

4. Discussion

This work shows that theory of mind is a double-edged sword. Through analysing in some detail pairs of RL agents endowed with theory of mind with different depths of mentalization in a mixed-motive game, we highlight how deeper mentalisation can protect agents, allowing them to act appropriately against deceptive partners. On the other hand, we also show how a higher level of theory of mind can be maladaptive when miscalibrated: Agents with deep recursive theory of mind, thinking three steps into the cognitive hierarchy, become skeptical against even random behavior and are trapped in a hypermentalised policy, believing they are surrounded by others that are out to get them.

Our work highlights how ToM-induced paranoia and its detrimental consequences are not only a function of the agent's own ToM but also of its environment and other agents. This is consistent with venerable observations (Simon, 1990; Bhui et al., 2021; Huys et al., 2015), and is relevant for the maladaptive behavior of machines (Schulz & Dayan, 2020). It also shows how complex phenomena like skepticism can arise even from optimal Bayesian inference (Bhui & Gershman, 2020; Alon et al., 2022) and how what an agent might think of as optimal Bayesian inference can go awry given confusion about the decision problem or an unfortunate environment (Huys et al., 2015)

Our work has particular relevance for computational psychiatry: Overly vigilant behavior is hypothesised as a generative factor in psychiatric symptoms, such as paranoia or anxiety (McLaren et al., 2022b; Sharp et al., 2011). Those suffering from these symptoms display hyper-mentalisation, inferring nefarious, complex intentions from sparse data. Our work offers a computational simulacrum of these phenomena, formalising recursive theory-of-mind and showing how paranoia can arise in purely reward-maximising, interactive agents with minor miscalibrations. Extreme, clinically relevant cases such as paranoia are likely to involve a combination of theory of mind going awry as well as inflexible priors on the general population (Barnby et al., 2022b;a). In our paradigm, these priors may be operationalized by holding biased beliefs that a partner is non-random. However, a singular focus on biased priors (that a partner is non-random) within an explanatory model would represent an implicit belief that others are more sophisticated than they appear. Such an explanation would still need to account for how this view of the world became reified. Our simulations take the first steps in measuring this development in a more tractable manner, offering crisp, testable, predictions.

Our simulations rely on a well-established game, and relatively simplistic agents. We did this to focus on exemplar emergent behaviors, but this naturally introduces limitations. First, this model assumes a strict k-level model. Future work might do away with a k-level model in favour of a more fluid DoM that allows adaptation to a partner (Camerer et al., 2004); this is an important direction for the future. A plausible prediction is that after learning a partner is not attempting to deceive, one's own DoM levels might reduce to fit the context (although the potential sophistication of the agent remains constant). This attenuation may be slower for those with high versus low psychiatric symptoms. In other cases, agents might use DoM as an intentional variable – something that they manipulate explicitly as part of utility maximization. Second, as we show, our skeptical DoM(2) agents are stuck in their overly defensive behavior. This is because they know they are powerless to cajole the DoM(1) into 'taking off its mask' and increasing its offers. This might be alleviated by imbuing receivers with thresholds of their own, which the DoM(2) might then be able to use to devise alternative strategies, such as manipulation of the DoM(1) into believing that its threshold is higher than is (much like the DoM(1) belief manipulation schema). Last, we use simple, fixed thresholds to determine the utility type of the sender. Replacing these egocentric utilities with social orientation utilities, like the Fehr-Schmidt (Fehr & Schmidt, 1999) inequality aversion, may yield other non-trivial effects of hypermentalization.

Another naturalistic extension of our model may also incorporate sophistication detection: the ability for an agent to recognise when it is up against a more sophisticated partner, even if it cannot change its own mentalisation depth. This is relevant in a number of real-world scenarios. For example, humans, particularly those who are paranoid, can believe that they are being confronted with agents who are in fact smarter than them and so whose actions lack a transparent rationale – one can sense a plot is afoot but not be able to fully conceptualise it. Such an extension would allow an agent to make heuristic responses, such as threats to exit a context, if they could not out-manoeuvre their partner strategically by increasing their mentalisation depth (Hertwig & Engel, 2016; Hula et al., 2018). A necessity of this modification requires a metacognitive understanding of the limitations of one's social cognition. Such metacognition might also be employed to make other decisions prior to drastic action, e.g., gathering more information about opponents (Schulz et al., 2023).

Finally, our work has relevance to Large Language Models (LLMs). Recent work has debated whether LLMs possess a theory of mind (Sap et al., 2023; Ullman, 2023; Kosinski, 2023) and preliminary work has investigated how LLMs might use its notional theory of mind in iterated games (Akata et al., 2023). This work has shown that LLMs can make such inferences, although strictly at inferences consistent with DoM(0). Our work suggests the merits of investigating deeper DoM in the context of LLMs: This is both for its deceptive potential and to avoid potentially misplaced paranoia in and between humans and artificial systems.

Acknowledgements

This research has been partly funded by Israel Science Foundation grant #1340/18 (NA), 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. *arXiv preprint*, 2023.

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.

Barnby, J. M., Mehta, M. A., and Moutoussis, M. The computational relationship between reinforcement learning, social inference, and paranoia. *PLoS computational biology*, 18(7):e1010326, 2022a.

Barnby, J. M., Raihani, N., and Dayan, P. Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, 225:105098, 2022b.

Barnby, J. M., Dayan, P., and Bell, V. Formalising social representation to explain psychiatric symptoms. *Trends in Cognitive Sciences*, 2023.

Barnett, S. A., Griffiths, T. L., and Hawkins, R. D. A pragmatic account of the weak evidence effect. *Open Mind*, pp. 1–14.

Bhui, R. and Gershman, S. J. Paradoxical effects of persuasive messages. *Decision*, 7(4):239–258, 2020. ISSN 23259973. doi: 10.1037/dec0000123.

Bhui, R., Lai, L., and Gershman, S. J. Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41:15–21, October 2021. ISSN 2352-1546. doi: 10.1016/j.cobeha.2021.02.015. URL https://www.sciencedirect.com/science/article/pii/S2352154621000371.

Camerer, C. F., Ho, T.-H., and Chong, J.-k. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., and Montague, P. Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, 57(3):463–473, 2008. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2007.12.020. URL https://www.sciencedirect.com/science/article/pii/S0896627307010331.

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.

Dennett, D. C. *The intentional stance*. MIT press, 1989.

Devaine, M., Hollard, G., and Daunizeau, J. Theory of mind: Did evolution fool us? *PLOS ONE*, 9(2):1–12, 02 2014a. doi: 10.1371/journal.pone.0087619. URL https://doi.org/10.1371/journal.pone.0087619.

Devaine, M., Hollard, G., and Daunizeau, J. Theory of Mind: Did Evolution Fool Us? *PLOS ONE*, 9(2):e87619, February 2014b. ISSN 1932-6203. doi: 10.1371/journal.pone.0087619. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087619. Publisher: Public Library of Science.

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.

Fehr, E. and Schmidt, K. M. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/2586885.

Freeman, D. Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7):685–692, 2016.

Frith, U. and Happé, F. Autism: beyond "theory of mind". *Cognition*, 50(1):115–132, 1994. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(94)90024-8. URL https://www.sciencedirect.com/science/article/pii/0010027794900248.

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.

Goodman, N. D. and Frank, M. C. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829, November 2016. ISSN 1364-6613. doi: 10.1016/j.tics.2016.08.005. URL https://www.sciencedirect.com/science/article/pii/S136466131630122X.

Harsanyi, J. C. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967.

Hertwig, R. and Engel, C. Homo Ignorans: Deliberately Choosing Not to Know. *Perspectives on Psychological Science*, 11(3):359–372, 2016. ISSN 17456924. doi: 10.1177/1745691616635594.

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 15537358. doi: 10.1371/journal.pcbi.1004254.

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., and Montague, P. R. A model of risk and mental state shifts during social interaction. *PLOS Computational Biology*, 14(2):1–20, 02 2018. doi: 10.1371/journal.pcbi.1005935. URL https://doi.org/10.1371/journal.pcbi.1005935.

Huys, Q. J., Guitart-Masip, M., Dolan, R. J., and Dayan, P. Decision-Theoretic Psychiatry. *Clinical Psychological Science*, 3(3):400–421, May 2015. ISSN 21677034. doi: 10.1177/2167702614562040. Publisher: SAGE Publications Inc.

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.

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].

Le, M., Boureau, Y. L., and Nickel, M. Revisiting the evaluation of theory of mind through question answering. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 5872–5877, 2019. doi: 10.18653/v1/d19-1598. ISBN: 9781950737901.

McLaren, V., Gallagher, M., Hopwood, C. J., and Sharp, C. Hypermentalizing and borderline personality disorder: A meta-analytic review. *American Journal of Psychotherapy*, 75(1):21–31, 2022a. doi: 10.1176/appi.psychotherapy.20210018. PMID: 35099264.

McLaren, V., Gallagher, M., Hopwood, C. J., and Sharp, C. Hypermentalizing and borderline personality disorder: A

meta-analytic review. *American Journal of Psychotherapy*, 75(1):21–31, 2022b.

Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Oey, L. A., Schachner, A., and Vul, E. Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 152(2):346, 2023.

O'Grady, C., Kliesch, C., Smith, K., and Scott-Phillips, T. C. The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, 36(4):313–322, 2015.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526, December 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/doesthe-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0. Publisher: Cambridge University Press.

Ransom, K., Voorspoels, W., Navarro, D., and Perfors, A. Where the truth lies: how sampling implications drive deception without lying. 2019.

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.

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, E. and Dayan, P. Computational Psychiatry for Computers. *Iscience*, 23(12):101772, 2020.

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.

Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., and Fonagy, P. Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(6):563–573, 2011.

Silver, D. and Veness, J. Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, 23, 2010.

Simon, H. A. Invariants of Human Behavior. *Annual review of psychology*, 41(1), 1990.

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].

Yoshida, W., Dziobek, I., Kliemann, D., Heekeren, H. R., Friston, K. J., and Dolan, R. J. Cooperation and heterogeneity of the autistic mind. *Journal of Neuroscience*, 30 (26):8815–8818, 2010.