Overly deep hierarchical mentalizing produces paranoia: a new formal theory

Alon, Nitay[1,2], Schulz, Lion[2], Bell, Vaughan[5], Moutoussis, Michael[6], Dayan, Peter[2,3], and Barnby, Joseph M.[4,7,8]

[1]*Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel*

[2]*Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

[3]*Department of Computer Science, University of Tübingen, Tübingen, Germany*

[4]*Department of Psychology, Royal Holloway University of London, London, UK*

[5]*Clinical, Educational, and Health Psychology, University College London, UK*

[6]*Department of Imaging Neuroscience, University College London, London, UK*

[7]*School of Psychiatry and Clinical Neuroscience, The University of Western Australia, Australia*

[8]*Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK*

February 7, 2024

1 Abstract

Humans need to be on their toes when interacting with competitive others to avoid being duped. Too much caution out of context can, however, be detrimental and produce false beliefs of intended harm. Here, we offer a formal account of this phenomenon through the lens of Theory of Mind. We simulate agents of different depths of mentalization within a simple game theoretic paradigm and show how, if aligned well, deep recursive mentalization gives rise to both successful deception as well as reasonable skepticism. However, we also show that if a self is mentalising too deeply - hyper-mentalising - false beliefs arise that a partner is trying to trick them maliciously, resulting in a material loss to the self. This theory offers a potential cognitive mechanism for suspiciousness,

paranoia, and conspiratorial ideation. Rather than a deficit in Theory of Mind, paranoia may arise from the application of overly strategic thinking to ingenuous behaviour.

2 Introduction

To be strategic, and thus sometimes also deceptive, we need to take into account the beliefs, desires and intentions of others. The cognitive process underlying such behaviour is theory of mind (ToM) - an agent's ability to reason about latent characteristics of others; what they know, want or plan (1; 2).

Signatures of ToM have captured the attention of computational scientists who have formalised ToM as a collection of social processes that enable inference and representation about the dynamic interaction between a self and other(s) (3; 4; 5; 6; 7; 8). At the most shallow level, an agent ('the self') simply considers the utility function (the desires) or beliefs of another agent (the 'other') based on their past behaviour (9; 10). This can be extended to deeper levels recursively: You can think about what I think you think I think (what you think, etc.). Hierarchical ToM - the ability to hold nested beliefs of ourselves and others (11; 12) - has been suggested as supporting the way that humans choose what to say or teach to maximise interpretability (13; 14), and as underlying cognition in social, competitive settings (15). It allows agents to hide information from others strategically, and to use an opponent's inference process against them in forms of deception, skepticism, and strategies to overcome these (16; 17).

With ToM's outsized role in human interaction (18), it is unsurprising that failures of ToM have been suggested as being at least part of the basis of several psychiatric disorders (19), such as autism (20; 21; 22), psychosis (23; 24; 25), and personality disorders (26; 27; 28; 29; 30; 31).

In patients with persecutory delusions and those with high paranoia, there is a tendency to make personal, external attributions – that is, explaining the causes of negative events through the malicious intentions of others (32). In borderline personality disorder (BPD), individuals are theorised as attributing an excessively high level of intentionality to sparse social data (26). Here, over-mentalising or hyper-mentalising is defined as "making excessively convoluted inferences based on others' social cues" and (33)) has been suggested as giving rise to paranoia in BPD (19),

and was shown empirically to be related to early stages of disorder (28). In both psychosis and BPD, and even more commonly in conspiratorial ideation (34), there is a higher risk of *over*-interpreting behaviour as being more sophisticated, intentional, and malicious.

Nevertheless, the cognitive mechanisms of this approach to paranoia within persecutory delusions and BPD have been hard to pin down and specify with the dynamic interaction and representation of social agents making mechanisms harder to examine (25; 35). There has been little work examining the role of cognitive recursion applied to BPD, paranoia, and persecutory delusions, aside from some notable exceptions (27; 36), which did not focus on false belief generation or maintenance.

Here, we offer an example of the ramifications of being adaptively and maladaptively strategic at different recursive levels. We use simulations based on Interactive Partially Observable Markov Decision Processes (IPOMDPs; 37) to suggest how this can help explain social cognitive processes that result in paranoia, suspiciousness, and/or conspiratorial ideation. We show how the degree of reasoning about the intentions of others (2; 18; 38) can be a protective factor against exploitation. However, we also demonstrate how this can go grossly awry: Selves that over-interpret actions of others make misplaced inferences about the others' strategic and deceptive intentions, with a malign effect on the reward garnered by the self.

We begin by emphasizing the importance of hierarchical ToM in mixed-motive games. These sequential social dilemmas (SSD) serve as a tractable testbed to observe the emergence of complex behaviour (16), as agents need to balance their reputation with material gains and losses. This work reinforces previous findings showing that agents with deep mental recursion, known as their Depth of Mentalisation (DoM) can successfully manipulate the beliefs of those with one-step lower DoM.

Next, we present the potential downside associated with maladaptively high DoM, i.e., hypermentalising. This pitfall is illustrated through a sequence of interactions between agents with mismatched DoM. We show that agents with maladaptively high DoM overestimate the complexity of their counterparts and overreact to sincere agents. This overreaction yields detrimental results. We then discuss how these results have the potential to explain some key aspects of

psychopathology.

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 paranoid beliefs and behaviour, and a possible mechanism underlying excessive recursive belief formation in general psychopathology. We show the AI community how ToM needs careful calibration to avoid counterproductive inference, and hence loss of credence and reward between agents. As a result, our work has key implications for AI safety and human-computer interaction.[1]

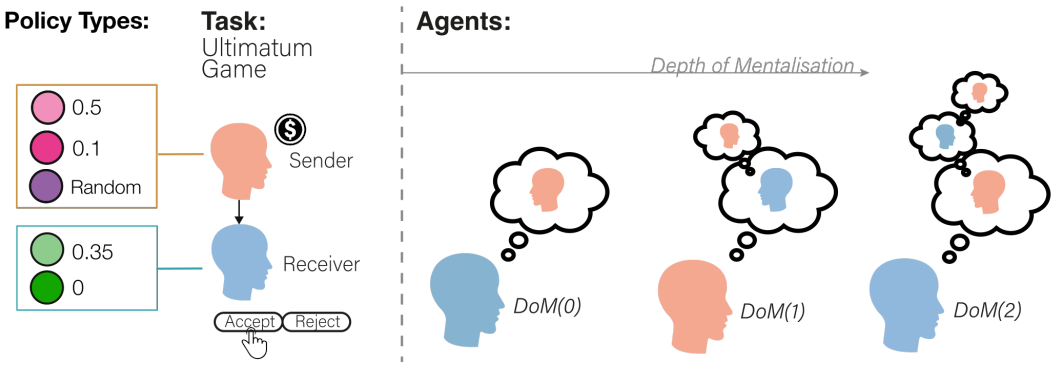

Figure 1: **Task and Agent Summary:** In the Ultimatum Game, a sender (orange) chooses how much of an endowment to send to a receiver (blue). The receiver then has a chance to either accept or reject this offer. If the receiver accepts, they both get to keep their portion of the endowment. If the receiver rejects, neither gets anything. 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. In addition, we introduce a random sender, sending uniformly distributed offers. The other type of sender and both receivers are endowed with Theory of Mind along with DoM $\in \{0, 1, 2\}$. This enables these agents to model their partners recursively, to a strictly limited extent. Both agents are characterized by their DoM level and by a threshold, representing in principle the minimal reward they are willing to accept

3 Paradigm and model

Mixed motive games offer a particularly useful test-bed to examine the rise of complex behaviour and test the role of opponent perception in social interactions. Generally speaking, a mixed-motive game is an interaction between two or more agents where there is mixed-desires in each agent's

[1]The code for this paper is available here

preferences over the outcome. One such game is the Prisoner's Dilemma, where both parties gain more from mutual cooperation than from mutual defection, while one side can gain an even higher reward by defecting from a cooperating partner. In this work, we match agents with increasing degrees of DoM in the Iterated Ultimatum game IUG (39)(Figure 1). This game is comprised of $T > 0$ repetitions (here $T = 12$) of the following game: a *sender*, S, is endowed with a some number of monetary units, set in this work to 1. They then offer the *receiver*, R, a partition of this endowment: the receiver would get a_S while the sender would get to keep $1 - a_S$ for themselves. The receiver then decides whether to accept the offer ($a_R = 1$) or to reject it ($a_R = 0$). In the latter case both parties get zero reward. The structure of the utilities makes the IUG a mixed-motive game: the sender's utility decreases with the offer size and so they are incentivised to offer the receiver less. However, if the sender offers too little, they will end up with nothing. Hence the sender has to balance their desires with those of the receiver to maximize their long-term utility.

We use the superscript t to denote the actions of both agents at trial $t \in [1, T]$: a_S^t, a_R^t. In turn, we define the *history* at time t as the sequence of offers and responses: $h^t = \langle a_S^1, a_R^1, \ldots, a_S^t, a_R^t \rangle$.

Apart from a particularly simple, random, sender, each agent, $i \in [R, S]$ is characterized by two parameters: its utility function: u_S, u_R and its DoM level: $k \in \{-1, 0, 1, 2\}$. The utility is governed by a threshold η, representing the minimal amount of money an agent is willing to receive. This allows us a simple control for testing how DoM interacts with utility preferences. In addition, thresholds serve as simple social orientation functions – those with higher thresholds are less likely to make compromises compared to those with low (or zero) thresholds. This serves to introduce diversity in the decision making process of the agents, representing in principle economic rational agents (reward maximizing agents, who act solely to maximize their utility and do not gain utility from other sources such as manipulation of others, social influence etc.). Other social orientation functions are expected to yield a different behaviour. For example, the Fehr-Schmidt utility (40) adds to the agent's utility gain (loss) from inequality aversion. We keep this option for future research.

In this work we consider two sender thresholds: $\eta_S \in \{0.1, 0.5\}$ and two receiver thresholds:

$\eta_R \in \{0.0, 0.35\}$. Formally, the utilities of agents with thresholds η_S, η_R are:

$$u_S^t(\eta_S, a_S^t, a_R^t) = (1 - a_S^t - \eta_S) * a_R^t \tag{1}$$

$$u_R^t(\eta_R, a_S^t, a_R^t) = (a_S^t - \eta_R) * a_R^t \tag{2}$$

Both agents seek to maximize their discounted long-term reward: $\sum_{t=1}^{T} u_i^t e^{(t-1 \log \gamma)}$, with a discount parameter $\gamma > 0$, here set to $\gamma = 0.99$.

Each agent (i) uses its DoM level (k) to compute the Q-values, $Q_{i=k}(a_i^t | h^{t-1}, \theta_i)$, which are used for action sampling (policy), π. We assume that both parties play a SoftMax policy, with a known temperature $\mathcal{T}$ (set in this work to $\mathcal{T} = 0.05$):

$$P_{i=k}^t(a_i^t | h^{t-1}, \theta_i) \propto \exp \frac{Q_{i=k}(a_i^t | h^{t-1}, \theta_i)}{\mathcal{T}} \tag{3}$$

The action's Q-value are computed as a function of the history and the agent's DoM level as described next.

We model the agents using the IPOMDP framework (37). This framework augments the POMDP model to account for modeling others. These models, denoted by θ, include all aspects of the other agent's decision-making characteristics and beliefs. In this task, these aspects include the other agent's threshold, but it may also include the other's agent's beliefs, including the beliefs of others about the self (i.e., nested beliefs). The level of recursion defines the agent's DoM level. In this work, we consider an iterated DoM level (41) - senders and receivers have odd- and even-numbered DoM respectively.

At the core of this framework are the DoM(-1) agents, also known as subintentional agents. These agents are characterized by lacking an opponent model, and are typically considered to be model-free RL agents. In this task, we consider *random* and the *threshold* DoM(-1) senders. The random sender makes offers uniformly random and does not adapt its behaviour to the receiver's response. We include this sender to examine the strategies used by senders and receivers to exploit the possible existence of a random other, since agents may mistake randomness for intentional or non-random behaviour.

The threshold DoM(-1) senders follow a reactive and myopic policy. If their current offer is accepted, they will offer less in the following iteration as they infer this acceptance as a sign that the offer was "too generous". On the other hand, if the offer is rejected, they will increase the next offer. Formally, these agents maintain a lower and upper bound representing the range of offers to consider:

$$L^t = L^{t-1} \cdot a_R^{t-1} + a_S^{t-1} \cdot (1 - a_R^{t-1}) \tag{4}$$

$$U^t = U^{t-1} \cdot (1 - a_R^{t-1}) + a_S^{t-1} \cdot (a_R^{t-1}) \tag{5}$$

with $L^0 = 0$ and $U^0 = 1$. In turn, these senders' Q-values are simply the utility from every action in the range $a_S^t \in [L^t, U^t]$:

$$Q^t_{S=-1}(a_S^t; \eta_S) = u_S^t(a_S^t, \eta_S) \tag{6}$$

The DoM(0) receiver models the sender as a DoM(-1) sender. In turn, it forms a belief about the type of the sender - either a random or a threshold sender: $\hat{\theta}_{S=-1} \in \{Random, 0.1, 0.5\}$. These beliefs are updated using IRL (9) and the nested models. Upon observing an offer a_S^t, the DoM(0) receiver computes the likelihood of the offer for each possible sender type and reweights them with its current beliefs:

$$b^t_{R=0}(\hat{\theta}_{S=-1}) = p^t_{R=0}(\hat{\theta}_{S=-1}|h^{t-1}, a_S^t) \propto P^t_{S=-1}(a_S^t|h^{t-1}, \hat{\theta}_{S=-1}) b^{t-1}_{R=0}(\hat{\theta}_{S=-1}) \tag{7}$$

where, $P^t_{S=-1}(a_S^t|h^{t-1}, \hat{\theta})$ is computed using the DoM(0) receiver's nested DoM(-1) sender model. We assume that the prior beliefs are both common knowledge and flat, making the updated belief common knowledge, as it is a deterministic function of the history, and the actions are fully observed. The DoM(0) receiver's Q-values are a combination of its immediate utility and the

discounted expected utility, given that it played a_R^t:

$$Q^t_{R=0}(a^t_R; \eta_R, b^t_{R=0}(\theta_{S=-1})) = \tag{8}$$
$$E_{a_S^{t+1} \sim \pi^*_{S=-1}}\left[u^t_R(a^t_S, \eta_R) \cdot a^t_R + \gamma \max_{a_R^{t+1}} \{Q^{t+1}_{R=0}(a^{t+1}_R; \eta_R, b^{t+1}_{R=0}(\theta_{S=-1}))\}\right]$$

where $E_{a_S^{t+1} \sim \pi^*_{S=-1}}$ is the expected future offer, weighted by the current belief.

Interacting with the simple DoM(-1) sender, these agents solve the optimal policy computation using the ExpectiMax algorithm (42). This planning algorithm computes the Q-value when playing against a stochastic adversary, by averaging over its expected actions.

Playing with the DoM(0) receiver in mind, the DoM(1) sender's belief includes inferences about the receiver's threshold and the receiver's beliefs about the sender's type. Due to the known priors and full observability, these nested beliefs are known to the DoM(1), but we specify them here for illustrative purposes:

$$b^t_{S=1}(\hat{\theta}_{R=0}, b^{t-1}_{R=0}(\hat{\theta}_{S=-1})) = \tag{9}$$
$$p^t_{S=1}(\hat{\theta}_{R=0}, b^{t-1}_{R=0} | h^{t-1}) \propto P^t_{R=0}(a^{t-1}_R | h^{t-2}, a^{t-1}_S, \hat{\theta}_{R=0}, b^{t-1}_{R=0}(\hat{\theta}_{S=-1})) b^{t-1}_{S=1}(\hat{\theta}_{R=0}, b^{t-2}_{R=0}(\hat{\theta}_{S=-1}))$$

The DoM(1) Q-values follow the same structure as the DoM(0) Q-values (Equation 8), where the expectation includes the updated beliefs of the DoM(0) receiver upon observing the offer:

$$Q^t_{S=1}(a^t_S; \eta_S, b^t_{S=1}(\hat{\theta}_{R=0}, b^{t-1}_{R=0})) = \tag{10}$$
$$E_{a_R^t \sim \pi^*_{R=0}}\left[u^t_S(a^t_S, \eta_S) \cdot a^t_R + \gamma \max_{a_S^{t+1}} \{Q^{t+1}_{S=1}(a^{t+1}_S; \eta_S, b^{t+1}_{S=1}(\hat{\theta}_{R=0}, b^t_{R=0}))\}\right]$$

Much like the DoM(0) receiver, the DoM(1) sender also infers through simulation how its actions affect the receiver's future actions. However, while the DoM(0) receiver manipulates the bounds of the DoM(-1) sender, the DoM(1) sender manipulates the beliefs of the DoM(0) receiver, in both cases to their own favour.

We also consider a DoM(2) receiver. This agent models the threshold sender as a DoM(1) sender and their belief update includes in addition to the type inference, the nested beliefs of

the DoM(1). These are the beliefs the DoM(1) sender ascribes to the presumed DoM(0) receiver. Notably, these receiver's also consider the random sender in its model. The belief update and Q-values computation follow the same formulation as in Equations (9, 10).

The DoM(1) and DoM(2) agents compute their Q-values using the IPOMCP planning algorithm (41), an extension of the POMCP algorithm to IPOMDP. Using their nested opponent model, these agents plan how to manipulate the policy of the DoM(0) receiver.

4 Results

We begin by analyzing the cases where the agents' DoM levels are typically matched, *i.e.*, where one agent has DoM$(k+1)$ and the other DoM(k). These simulations establish a baseline of typical strategic behaviour stemming from the higher DoM agent's ability to manipulate its counterpart.

First, since savvy opponents should act deceptively, a high DoM agent can confuse naive behaviour as arising instead from clever malevolence. We account for this by analyzing the counter-deceptive reasoning applied by high DoM agents and show how this skepticism harms them when matched with benign agents.

Second, high DoM agents believe themselves to be interacting with moderately high DoM partners, and can believe that such moderately high DoM partners are strategically impervious to the responses of what they assume to be their even lower DoM opponents. Thus, the high DoM agents can exhibit a form of helplessness when playing low DoM partners, when in fact they would be perfectly capable of exploiting them appropriately. We illustrate this by simulating the DoM(2) receiver and the DoM(-1) sender. Due to the strictness of opponent reasoning of the cognitive hierarchy, the DoM(2) models its counterpart as DoM(1), misinterpreting the behaviour of the DoM(-1) sender.

From these simple reward-maximising mechanisms enacted in a competitive, interactive context, we conclude that the combination of a deceptive opponent and a maladaptively high level of DoM gives rise to over-mentalising and loss of reward.

4.1 Baseline behaviour

Theory of Mind (ToM) is used for both inference and planning. For example, when the DoM(0) receiver observes an offer by the sender, its belief update allows it to identify the type of sender by inverting the offer to infer the sender's characteristics. Here, the sender is assumed by the receiver to have a DoM(-1) policy: it is unable to mentalise about the receiver and simply to hold to its policy. The DoM(0) receiver can then use its model of the sender to simulate how each sender type would respond to the receiver's action, weighing the optimal response according to its beliefs. Thus, it can manipulate the sender's behaviour to its benefit, up to a irreducible uncertainty. This example serves to show that those with higher DoM can theoretically gain an advantage over those with lower DoM.

4.1.1 DoM(-1) sender and DoM(0) Receiver: Näive utility calculus

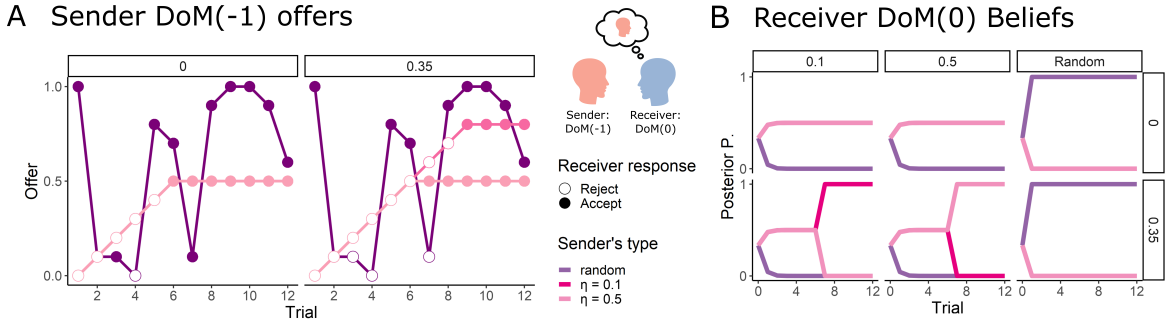

Figure 2: **Illustration of DoM(0) IRL: (A,B)** In interacting with the DoM(-1) sender **(A)**, the DoM(0) receiver makes inferences about the sender's type **(B)**. Notably, the first offer is usually sufficient to tell the random sender from the threshold senders. When the receiver's belief favours the threshold sender, the receiver manipulates the sender by rejecting the offers until a desired offer is met, according to the receiver's threshold.

Following the properties of hierarchical mentalising, we begin with the first dyad of adaptively aligned DoM levels - a DoM(-1) sender interacting with a DoM(0) receiver. The DoM(0) inference about the DoM(-1) type is displayed in Figure 2(B). Crucially, the first offer here suffices to tell the random sender from the threshold ones (since it is so high). After making this distinction, the receiver adapts its policy. If the beliefs support the threshold sender, the optimal policy is to reject the offers strategically, pushing the lower bound upward until a desired level is met. Figure 2(A)

shows this manipulation as a function of the receiver's threshold – the zero threshold receiver's acceptable offer is 0.5 (which is the maximal offer the DoM(-1) sender with $\eta = 0.5$ is willing to make), while the 0.35 threshold receiver is "demanding" for a higher offer to maximize its long term cumulative reward. On the other hand, if the DoM(0) receiver believes it is facing the random sender, it accepts any offer that satisfies its threshold, as the random agent cannot be manipulated. This is appropriate and adaptive to the context.

4.1.2 DoM(1) sender and DoM(0) Receiver: Deception through induced false beliefs

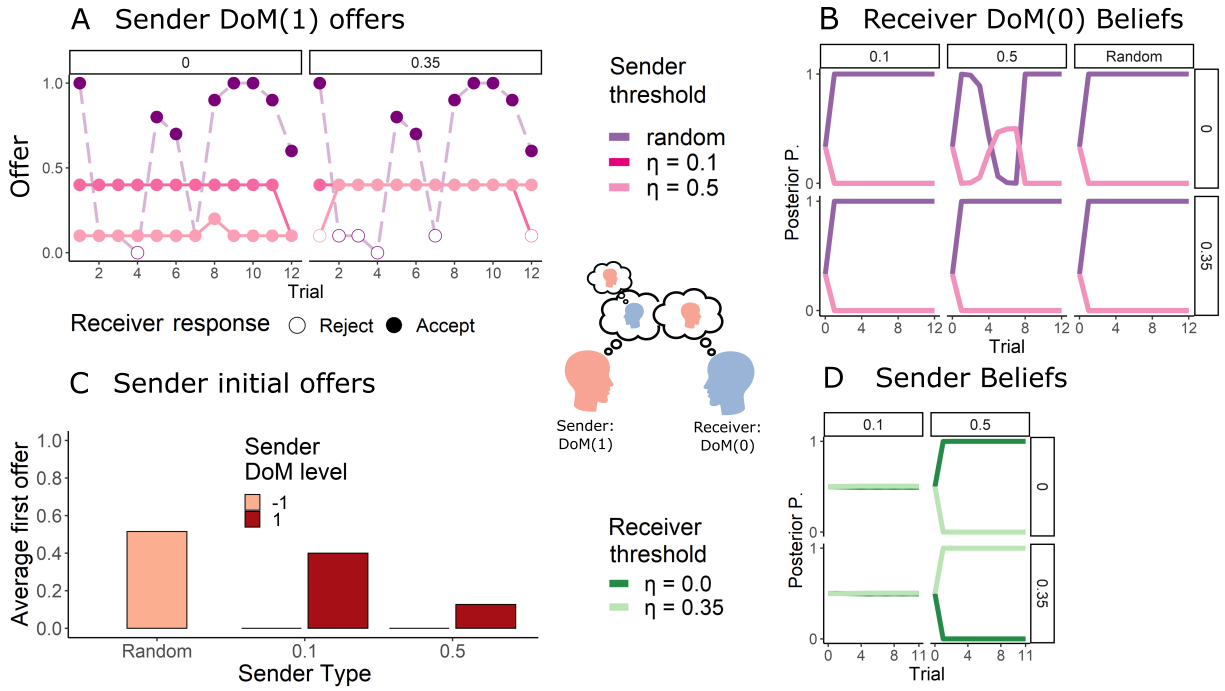

Figure 3: **Dynamics of the DoM(1) manipulation:** The DoM(1) offers are strategically aimed at shifting the DoM(0) belief in favour of the random sender hypothesis. This strategy naturally arises out of simple reward-maximising agents in a mixed motive setting. **(A)** The sender's initial offer "mimics" that of the random sender, before subsequently defecting. **(B)**, The DoM(0) receiver beliefs are completely hijacked by the DoM(1) deception. The DoM(1) offers are deliberately high, to be classified by the DoM(0) beliefs as coming from a random sender. While the DoM(-1) sender's first offer is 0.0, the DoM(1) sends between 0.1 and 0.4.**(C)** The DoM(1) sender's deception is characterized by making a relatively high first offer. This offer is highly atypical for a DoM(-1) threshold sender. **(D)** using the same IRL concept, the DoM(1) makes inferences about the DoM(0) receiver's type from its responses

The DoM(1) sender uses its DoM(0) nested model for optimal policy computation. It emulates the DoM(0) inference process, and consequently predicts its policy. Given the policies depicted in Figure 2, the DoM(1) sender's policy is to masquerade itself as a random DoM(-1) agent, causing the DoM(0) to accept any offer (respecting the receiver's threshold). Thus, the DoM(1) sender could avoid the strategic rejection of the DoM(0) receiver. This is achieved by making offers that are highly unlikely for a threshold sender (from the DoM(0) receiver's perspective), illustrated in Figure 3(C). Given the low SoftMax temperature, the DoM(1) correctly infers that the DoM(0) receiver would infer any offer other than 0.0 as typical behaviour of the random sender, as the DoM(-1) is expected to start offering nothing.

Following this ploy, the DoM(1) sender repeatedly sends the bare minimal offer (presented in 3(A)), to extract access wealth at the expense of the DoM(0). This deception utilizes a pitfall of the DoM(0) inference process – the likelihood of any action is the same under the random sender generative model hypothesis. Thus, the likelihood of a flat trajectory of offers is the same as the likelihood of the "true" random offers - making the DoM(0) receiver unable to tell the true random from the fake one as depicted in Figure 3(B).

4.1.3 DoM(1) sender and DoM(2) Receiver: Defying deception with deception

Being aware of this ploy, the DoM(2) receiver uses its nested model of the DoM(1) sender to flag seemingly "random" offers as having been generated by a savvy adversary, identifying the masquerader and reacting accordingly (Figure 4). Applying the same deceptive principles as the DoM(1), the DoM(2) acts in a way that causes the DoM(1) to believe falsely that it is matched with the higher $\eta = 0.35$ receiver, thus pressuring the sender to improve its offers. This yields a higher reward compared to the limited-opponent modelling DoM(0) receiver. Notably, due to the built-in advantage of the sender in this task (has to offer at most 0.4) the advantage of the DoM(2) is manifested in the decrease of the reward ratio compared to the DoM(0).

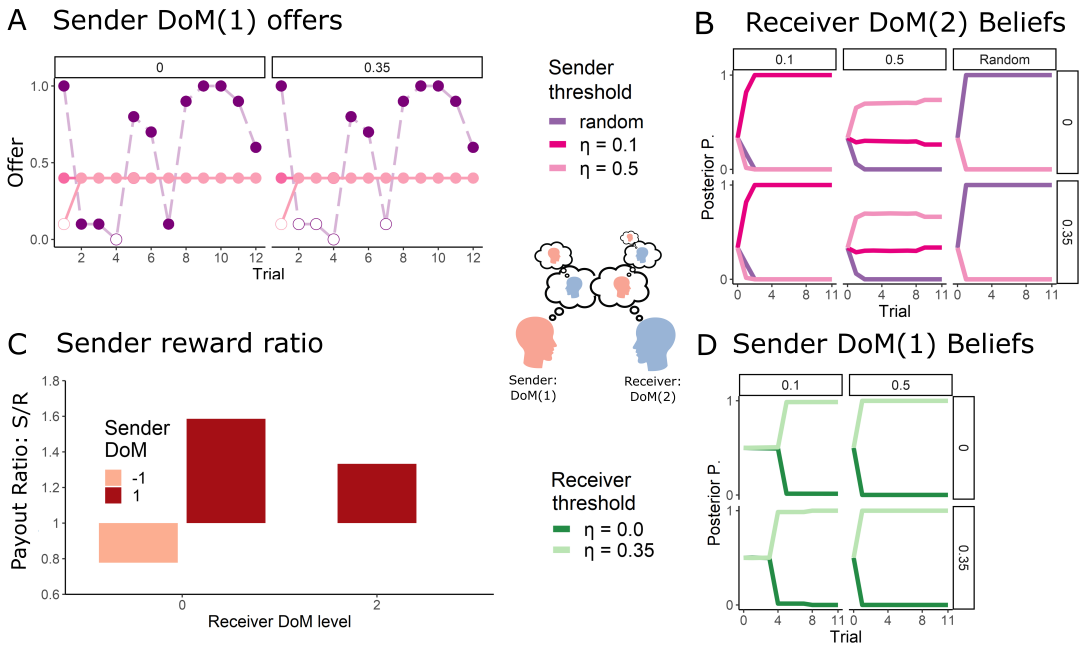

Figure 4: **Dynamics of the DoM(2) counter-manipulation: (A)** The DoM(2), masquerading as the high threshold receiver, rejects low offers. This encourages the DoM(1) sender with a high threshold to improve its offers, while having little effect on the already "generous" $\eta = 0.1$ DoM(1) sender. **(B)**, the DoM(2) receiver correctly reads the DoM(1) sender's bluff, while manipulating the latter's beliefs **(D)**. **(C)** Typically, the agent with the higher DoM gains a higher reward than the lower DoM agent. The y-axis measures the ratio between the receiver and sender's total reward. Due to the asymmetric nature of the IUG, the DoM(2) receiver superiority is manifested in its ability to lower the DoM(1) sender advantage.

We conclude that when appropriately matched, having DoM$(k+1)$ compared to a DoM(k) counterpart is beneficial. These findings reinforce previous studies highlighting the advantages of higher DoM in mixed-motive games. Figure 4(C) illustrates this supremacy - the total reward ratio is always in favour of the higher DoM agent.

4.2 Skepticism and paranoia in DoM(2)

While benefiting its holder, DoM is a double-edged sword. A mismatched DoM agent may misinterpret the actions of those with even lower DoM levels than they expect, misinterpreting simplistic behaviour as the product of Machiavellian sophistication. Here, knowing that its DoM(1) sender opponent would masquerade itself as a random sender, the DoM(2) receiver is susceptible to interpreting random behaviour as having been generated by the DoM(1) sender. This leads to

delayed detection of the random sender, as more "random" behaviour is required to confirm that the sender is genuinely random. The effect of the random-like ruse is that it takes, on average, 5 trials for the belief distribution of the DoM(2) receiver to converge to the random type compared to the 2 trials it takes on average for the DoM(0).

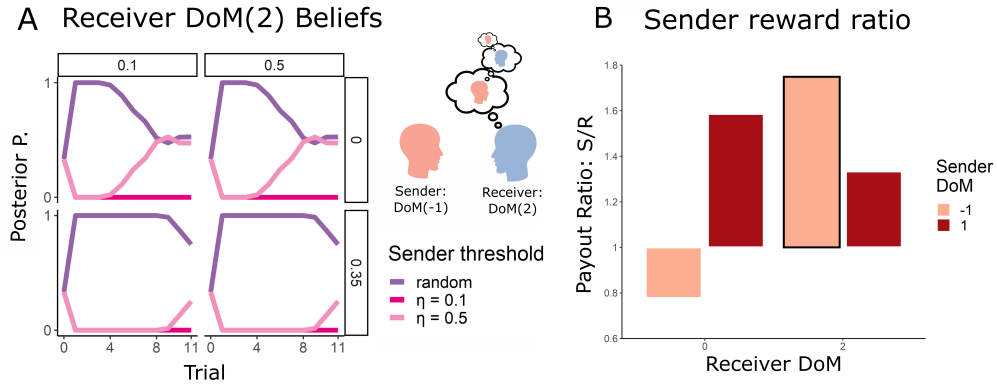

Figure 5: **Effects of maladaptive DoM**: **(A)** The offers of the DoM(-1) sender lie outside the DoM(2) opponent model and are viewed as coming from a random sender.**(B)** In turn, the receiver's docile policy means that they are willing to accept any offer, yielding them a low reward

While delayed random identification has a limited effect on the DoM(2) reward, the other type of maladaptive DoM, over-mentalising, yields a more severly detrimental outcome.

Over-attribution of intentionality means that the DoM(2) receiver fails to model the DoM(-1) sender properly. The low offers of the threshold DoM(-1) senders are atypical for the savvy, random-pretending DoM(1). Thus, the DoM(2) receiver's tendency to interpret the "random" behaviour of the DoM(1) as a sign of intentional strategy causes it to interpret any of the non-random DoM(-1) actions as a sign of random behaviour as depicted in Figure 5(A).

Given that the best response to a truly random sender is just to accept anything above one's threshold, the trapped DoM(2) receiver accepts most of the threshold DoM(-1) sender offers. However, these simplistic senders will improve their offers only if rejected, otherwise, they continue to make the same low offer. The detriment to the receiver is evident in Figure 5(B). In effect, the DoM(2) never acts to cause the DoM(-1) to show itself to be able to be changed, and so never encounters evidence against its own beliefs.

While the DoM(2) falls for the same ploy applied by the DoM(1) sender against the DoM(0) receiver, the causal mechanism between the two differs. The DoM(2) is a victim of its sophistication, and the incorrectly perceived sophistication of its partner, and not the victim of a truly savvy opponent.

5 Discussion

This work shows that hierarchical mentalising is a double-edged sword. We analysed pairs of RL agents endowed with ToM at different depths of mentalization in a mixed-motive game. When agents correctly model their opponent's degree of sophistication, they can protect themselves, acting appropriately against deceptive partners. These simulations are aligned with the hypothesis that ToM has evolved out of the need to survive and succeed in complex mixed-motive environments (43; 44; 45). On the other hand, we also show how high DoM can be maladaptive when miscalibrated: Agents thinking three steps into the cognitive hierarchy become sceptical against even random behaviour and are trapped in a hypermentalised policy, believing they are surrounded by sophisticated others that are out to trick them. This phenomenon, generated purely from two simple reward-maximising agents in an interactive context, makes for a plausible explanation for the generation and maintenance of psychopathological states, such as paranoia, where misperceiving others' negative intentions is a central feature and important source of disability.

Our work highlights how maladaptive DoM are functions of the agent's own beliefs, its environment and the beliefs of other agents. This is consistent with prior observations (46; 47; 48), and is relevant for the maladaptive behaviour of machines (49). It also shows how complex phenomena like scepticism can arise even from optimal Bayesian inference (16; 50) and how what an agent might think of as optimal Bayesian inference can go awry given the confusion about the decision problem or an unfortunate environment (48)

The over-attribution of negative social intentions is a central feature in paranoid delusions and borderline personality disorder (32) and hyper-mentalising has been identified as an important transdiagnostic feature in psychopathology more broadly (19; 26). Our work offers a computational model of these phenomena, formalising a theory of how hierarchical, recursive social cognition

gone awry may explain how paranoia can arise and be maintained in purely reward-maximising, interactive agents, even with minor miscalibrations. This cognitive mechanism may play a crucial role in the formation of persecutory delusions, along with inflexible priors about interaction partners (51; 52; 53), noisy mental models of others (54; 55), social hypersensitivity (56), and biased social values (57).

Our simulations rely on a well–established game, and relatively simplistic agents to focus on exemplar emergent behaviours explaining the production of false beliefs around the strategic and malicious nature of others, but this naturally introduces limitations. First, we use simple, fixed thresholds to determine the utility type of the sender. Indeed, Fehr-Schmidt (FS) (40) or FS-like utility functions are typically used to assess rejection in social contexts (e.g. (36; 57)), although we opted to remove this to isolate the effect of DoM. Replacing these egocentric utilities with social orientation utilities, like inequity aversion (58), may yield other non-trivial effects of hypermentalising.

Second, our model assumes a strict k-level model. This means that an agent's interpretation of the opponent is bounded to a fixed level of DoM, making the higher DoM agents susceptible to over-mentalisation and unable to assume otherwise. One remedy for this problem, which future work may explore, is adopting a mixture model view of the cognitive hierarchy. In this version, suggested by (11), a DoM(k) views the world as composed of different levels of DoM levels, ranging from $(k-1)$ to (-1), distributed according to a truncated Poisson distribution. This model may solve the problem of over-mentalisation, as the higher DoM agent no longer treats others as having a fixed DoM($k-1$) but rather has having multiple (unknown) DoM levels. However, it comes with an increase in the computational costs and complexity of the inference process. Nevertheless, from a highly paranoid perspective, we would predict that higher, more sophisticated DoM levels would be disproportionately inferred by those with BPD, persecutory delusions, or heightened paranoia compared to those with low paranoia.

Another future direction for solving fixed over-mentalisation is to make the DoM level an intentional, adaptive parameter. For example, after learning a partner is not attempting to deceive, one's own DoM might reduce to fit the context (although the potential sophistication of the agent

remains constant). A potential source and consequence of psychiatric symptoms might be a sloth in making this reduction even when the costs are high in terms of both computation and utility. Again, we predict that those with high vs low paranoia would enter into high DoM states much faster and take longer to reduce to adopting a lower DoM when the environment is evidently less competitive.

Another natural 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 DoM. This is relevant in several real-world scenarios and may offer a heuristic 'cheat' to the k-level hierarchy rationale. For example, humans, particularly those who are paranoid, can believe that they are being confronted with agents who are smarter than them and 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 (36; 59). 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 before drastic action, e.g., gathering more information about opponents (60).

6 Conflicts of Interest

None to declare

7 Funding

Funding was from 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".

8 Author Contributions

Nitay Alon: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Writing - original draft preparation. Lion Schulz: Formal Analysis, Visualisation Vaughan Bell: Writing - review and editing. Michael Moutoussis: Writing - review and editing. Peter Dayan: Conceptualisation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing - review and editing. Joseph M Barnby: Conceptualisation, Project Administration, Supervision, Visualisation, Writing - original draft preparation, Writing - review and editing.

References

[1] Dennett DC. The intentional stance. MIT press; 1989.

[2] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences. 1978 Dec;1(4):515-26. Publisher: Cambridge University Press. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0.

[3] Barnby JM, Bellucci G, Alon N, Schilbach L, Bell V, Frith CD, et al. Beyond Theory of Mind: A formal framework for social inference and representation. PsyArXiv. 2023. Available from: https://doi.org/10.31234/osf.io/cmgu7.

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

[5] Baker C, Saxe R, Tenenbaum J. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. Proceedings of the Annual Meeting of the Cognitive Science Society. 2011;33(33). Available from: https://escholarship.org/uc/item/5rk7z59q.

[6] Baker C, Tenenbaum J. Modeling Human Plan Recognition Using Bayesian Theory of Mind. Plan, Activity, and Intent Recognition: Theory and Practice. 2014 Mar:177-204.

[7] Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour. 2017 Mar;1(4):1-10. Number: 4 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41562-017-0064.

[8] Goodman ND, Baker CL, Bonawitz EB, Mansinghka VK, Gopnik A, Wellman H, et al. Intuitive theories of mind: A rational approach to false belief. In: Proceedings of the twenty-eighth annual conference of the cognitive science society. vol. 6. Cognitive Science Society Vancouver; 2006. .

[9] Ng AY, Russell S. Algorithms for Inverse Reinforcement Learning. In: in Proc. 17th International Conf. on Machine Learning. Morgan Kaufmann; 2000. p. 663-70.

[10] Jara-Ettinger J. Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences. 2019 Oct;29:105-10. Available from: https://www.sciencedirect.com/science/article/pii/S2352154618302055.

[11] Camerer CF, Ho TH, Chong Jk. A cognitive hierarchy model of games. The Quarterly Journal of Economics. 2004;119(3):861-98.

[12] O'Grady C, Kliesch C, Smith K, Scott-Phillips TC. The ease and extent of recursive mindreading, across implicit and explicit tasks. Evolution and Human Behavior. 2015;36(4):313-22.

[13] Goodman ND, Frank MC. Pragmatic Language Interpretation as Probabilistic Inference. Trends in Cognitive Sciences. 2016 Nov;20(11):818-29. Available from: https://www.sciencedirect.com/science/article/pii/S136466131630122X.

[14] Barnett SA, Griffiths TL, Hawkins RD. A pragmatic account of the weak evidence effect. Open Mind. 2022:1-14.

[15] Devaine M, Hollard G, Daunizeau J. The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? PLOS Computational Biology. 2014 Dec;10(12):e1003992. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003992.

[16] Alon N, Schulz L, Rosenschein JS, Dayan P. A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. Open Mind. 2023 08;7:608-24. Available from: https://doi.org/10.1162/opmi_a_00097.

[17] Doshi P, Qu X, Goodie A. Chapter 8 - Decision-Theoretic Planning in Multiagent Settings with Application to Behavioral Modeling. In: Sukthankar G, Geib C, Bui HH, Pynadath DV, Goldman RP, editors. Plan, Activity, and Intent Recognition. Boston: Morgan Kaufmann; 2014. p. 205-24. Available from: https://www.sciencedirect.com/science/article/pii/B9780123985323000087.

[18] Devaine M, Hollard G, Daunizeau J. Theory of Mind: Did Evolution Fool Us? PLOS ONE. 2014 02;9(2):1-12. Available from: https://doi.org/10.1371/journal.pone.0087619.

[19] McLaren V, Gallagher M, Hopwood CJ, Sharp C. Hypermentalizing and borderline personality disorder: A meta-analytic review. American Journal of Psychotherapy. 2022;75(1):21-31.

[20] Frith U, Happé F. Autism: beyond "theory of mind". Cognition. 1994;50(1):115-32. Available from: https://www.sciencedirect.com/science/article/pii/0010027794900248.

[21] Yoshida W, Dziobek I, Kliemann D, Heekeren HR, Friston KJ, Dolan RJ. Cooperation and heterogeneity of the autistic mind. Journal of Neuroscience. 2010;30(26):8815-8.

[22] Chiu PH, Kayali MA, Kishida KT, Tomlin D, Klinger LG, Klinger MR, et al. Self Responses along Cingulate Cortex Reveal Quantitative Neural Phenotype for High-Functioning Autism. Neuron. 2008;57(3):463-73. Available from: https://www.sciencedirect.com/science/article/pii/S0896627307010331.

[23] Bentall R, Kinderman P. Psychological processes and delusional beliefs: Implications for the treatment of paranoid states. Outcome and innovation in psychological treatment of schizophrenia. 1998.

[24] Randall F, Corcoran R, Day J, Bentall R. Attention, theory of mind, and causal attributions in people with persecutory delusions: A preliminary investigation. Cognitive neuropsychiatry. 2003;8(4):287-94.

[25] Penn DL, Sanna LJ, Roberts DL. Social cognition in schizophrenia: an overview. Schizophrenia bulletin. 2008;34(3):408-11.

[26] Sharp C, Pane H, Ha C, Venta A, Patel AB, Sturek J, et al. Theory of mind and emotion regulation difficulties in adolescents with borderline traits. Journal of the American Academy of Child & Adolescent Psychiatry. 2011;50(6):563-73.

[27] Hula A, Montague PR, Dayan P. Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange. PLoS Computational Biology. 2015;11(6):e1004254.

[28] Galvez-Merlin A, Lopez-Villatoro JM, de la Higuera-Gonzalez P, de la Torre-Luque A, Reneses-Prieto B, Diaz-Marsa M, et al. Social Cognition Deficits in Borderline Personality Disorder: Clinical Relevance. Psychiatry Research. 2023:115675.

[29] Rifkin-Zybutz R, Moran P, Nolte T, Feigenbaum J, King-Casas B, Personality L, et al. Impaired mentalizing in depression and the effects of borderline personality disorder on this relationship. Borderline personality disorder and emotion dysregulation. 2021;8(1):15.

[30] Euler S, Nolte T, Constantinou M, Griem J, Montague PR, Fonagy P, et al. Interpersonal problems in borderline personality disorder: associations with mentalizing, emotion regulation, and impulsiveness. Journal of Personality Disorders. 2021;35(2):177-93.

[31] King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR. The rupture and repair of cooperation in borderline personality disorder. science. 2008;321(5890):806-10.

[32] Buck B, Browne J, Gagen EC, Penn DL. Hostile attribution bias in schizophrenia-spectrum disorders: narrative review of the literature and persisting questions. Journal of Mental Health. 2023;32(1):132-49.

[33] Fonagy P, Luyten P, Bateman A. Translation: Mentalizing as treatment target in borderline personality disorder. Personality Disorders: Theory, Research, and Treatment. 2015;6(4):380.

[34] Bowes SM, Costello TH, Tasimi A. The conspiratorial mind: A meta-analytic review of motivational and personological correlates. Psychological Bulletin. 2023.

[35] Bell V, Mills KL, Modinos G, Wilkinson S. Rethinking social cognition in light of psychosis: reciprocal implications for cognition and psychopathology. Clinical Psychological Science. 2017;5(3):537-50.

[36] Hula A, Vilares I, Lohrenz T, Dayan P, Montague PR. A model of risk and mental state shifts during social interaction. PLOS Computational Biology. 2018 02;14(2):1-20. Available from: https://doi.org/10.1371/journal.pcbi.1005935.

[37] Gmytrasiewicz PJ, Doshi P. A Framework for Sequential Planning in Multi-Agent Settings. Journal of Artificial Intelligence Research. 2005 Jul;24:49-79. Available from: https://jair.org/index.php/jair/article/view/10414.

[38] Ho MK, Saxe R, Cushman F. Planning with Theory of Mind. Trends in Cognitive Sciences. 2022 Sep. Available from: https://www.sciencedirect.com/science/article/pii/S1364661322001851.

[39] Alon N, Schulz L, Dayan P, Barnby JM. Between prudence and paranoia: Theory of Mind gone right, and wrong. In: First Workshop on Theory of Mind in Communicating Agents; 2023. Available from: https://openreview.net/forum?id=gB9zrEjhZD.

[40] Fehr E, Schmidt KM. A Theory of Fairness, Competition, and Cooperation. The Quarterly Journal of Economics. 1999;114(3):817-68. Available from: http://www.jstor.org/stable/2586885.

[41] Hula A, Montague PR, Dayan P. Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange. PLOS Computational Biology. 2015 Jun;11(6):e1004254. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004254.

[42] Hutter M. Universal artificial intelligence: Sequential decisions based on algorithmic probability. Springer Science & Business Media; 2004.

[43] Lyons M, Caldwell T, Shultz S. Mind-reading and manipulation—Is Machiavellianism related to theory of mind? Journal of Evolutionary Psychology. 2010;8(3):261-74.

[44] Whiten A, Byrne RW. The Machiavellian intelligence hypotheses. 1988.

[45] Qi W, Vul E. Adaptive behavior in variable games requires theory of mind. 2023 Dec. Publisher: OSF. Available from: https://osf.io/7kw4z.

[46] Simon HA. Invariants of Human Behavior. Annual review of psychology. 1990;41(1).

[47] Bhui R, Lai L, Gershman SJ. Resource-rational decision making. Current Opinion in Behavioral Sciences. 2021 Oct;41:15-21. Available from: https://www.sciencedirect.com/science/article/pii/S2352154621000371.

[48] Huys QJM, Guitart-Masip M, Dolan RJ, Dayan P. Decision-Theoretic Psychiatry. Clinical Psychological Science. 2015 May;3(3):400-21. Publisher: SAGE Publications Inc.

[49] Schulz E, Dayan P. Computational Psychiatry for Computers. Iscience. 2020;23(12):101772.

[50] Bhui R, Gershman SJ. Paradoxical effects of persuasive messages. Decision. 2020;7(4):239-58.

[51] Barnby JM, Raihani N, Dayan P. Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. Cognition. 2022;225:105098.

[52] Diaconescu AO, Wellstein KV, Kasper L, Mathys C, Stephan KE. Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. Journal of Abnormal Psychology. 2020;129(6):556.

[53] Wellstein KV, Diaconescu AO, Bischof M, Rüesch A, Paolini G, Aponte EA, et al. Inflexible social inference in individuals with subclinical persecutory delusional tendencies. Schizophrenia Research. 2020;215:344-51.

[54] Adams RA, Vincent P, Benrimoh D, Friston KJ, Parr T. Everything is connected: inference and attractors in delusions. Schizophrenia research. 2022;245:5-22.

[55] Barnby JM, Mehta MA, Moutoussis M. The computational relationship between reinforcement learning, social inference, and paranoia. PLoS computational biology. 2022;18(7):e1010326.

[56] Henco L, Diaconescu AO, Lahnakoski JM, Brandi ML, Hörmann S, Hennings J, et al. Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. PLoS computational biology. 2020;16(9):e1008162.

[57] Kazinka R, Kwashie AN, Pratt D, Vilares I, MacDonald III AW. Value representations of spite sensitivity in psychosis on the Minnesota Trust Game. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2023.

[58] Hughes E, Leibo JZ, Phillips MG, Tuyls K, Duéñez-Guzmán EA, Castañeda AG, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. arXiv:180308884 [cs, q-bio]. 2018 Sep. ArXiv: 1803.08884. Available from: http://arxiv.org/abs/1803.08884.

[59] Hertwig R, Engel C. Homo Ignorans: Deliberately Choosing Not to Know. Perspectives on Psychological Science. 2016;11(3):359-72.

[60] Schulz L, Fleming SM, Dayan P. Metacognitive Computations for Information Search: Confidence in Control. Psychological Review. 2023. Available from: https://doi.org/10.1037/rev0000401.

[61] Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of the 20th international joint conference on Artifical intelligence. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2007. p. 2586-91.

[62] Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB. The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. Trends in Cognitive Sciences. 2016 Aug;20(8):589-604. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1364661316300535.

[63] Hjort NL, Holmes C, Müller P, Walker SG. Bayesian nonparametrics. vol. 28. Cambridge University Press; 2010.

9 Appendix

9.1 IRL as ToM inference

As presented above, inference in ToM can be seen as an extension of inverse RL (IRL; 9). Bayesian IRL (5; 61) requires an observer to make a Bayesian inference about the utility (reward) function of an agent from a sequence of observed behaviour $o^{0:T}$:

$$p(u|o^{0:T}) \propto P(o^{0:T}|u)p(u) \tag{11}$$

The DoM(0) inference follows this principle, inferring about the DoM(-1)'s utility function from its behaviour, using a nested model. This observation was made before (10), and framed as the Näive utility calculus (62). Formally, this inference requires a commonly known behaviour of the DoM(-1) (Equation 5). This behaviour is composed of the DoM(-1) Q-values (Equation 6) and its policy (Equation 3). Plugging into Equation 7 give rise to the IRL process, effectively a posterior distribution over the utility functions, as in Figure2B.

While following the same principles, namely inverting the behaviour to infer about the mechanism, higher DoM agents inference goes beyond utility inference. In this case, the inference also includes the agent's beliefs (Equation 9). Notably, if the common prior or action observability assumptions are revoked, the inference process yields a multi-dimensional distribution: $p(\theta \times b(\cdot)) = p(\theta) \times p(b(\cdot))$. The first component is similar to the utility inference of the DoM(0) agent, while the second one is a distribution over distributions (63). We refer the reader to (37) for a

full introduction of belief update in this case.