

The Role of Theory of Mind in Bayesian Mixed Motive Games

Thesis for the degree of “Doctor of Philosophy”

by

Nitay Alon

Department of Computer Science
The Hebrew University of Jerusalem

Submitted to the Senate of the Hebrew University of Jerusalem

10 2026

The Role of Theory of Mind in Bayesian Mixed Motive Games

Thesis for the degree of “Doctor of Philosophy”

by

Nitay Alon

Department of Computer Science
The Hebrew University of Jerusalem

Submitted to the Senate of the Hebrew University of Jerusalem

10 2026

This work was carried out under the supervision of Prof. Jeffrey S. Rosenschein and Prof. Peter Dayan

Acknowledgements

It has been said that it takes a village to raise a child. The same is true for raising a PhD student. This work represents my academic journey, an expedition which was accompanied by my advisors, my colleagues, my friends and my family — to each and every one of them I owe an endless gratitude.

First and foremost, my advisors — Jeff Rosenschein and Peter Dayan. Finding an optimal PhD advisor is an NP-Hard task, and I was fortunate to find two. To Jeff, the kindest person in AI as I fondly introduce him to people, who taught me endless lessons about how to transform my ideas into papers, how to mentor students, how to combine science and art (music). To Peter, the world's ultimate proof reader, who endured endless hours of fixing my hasty drafts, listening to my endless stream of ideas. The professor who introduced me to Theory of Mind and metastasized me into the Machiavellian Theory of Mind church. I cannot stress my gratitude to Jeff and Peter. I can only hope that I will live up to be half the mentor and advisor to my people as they were to me. I'd like to thank my PhD board - Prof. Jakob Foerster and Prof. Prashant Doshi, for their stream of ideas, critics and suggestions on how to improve my work.

My work is the fruit of joint labor. I was fortunate to find and collaborate with great minds, and if I accomplished anything, it was by standing on the shoulders of a giant. Joe Barnby — my endless collaborator, my mentor, my friend. Joe introduced me to Computational Psychiatry, to modeling complex human behavior using crisp models. We spend countless hours talking about ideas on how ToM is related to many aspects of social interactions, the existence of ToM in non human agents and so many ideas that I'm sorry I haven't kept a well organized record of them — each of which would make an amazing paper. To Stefan Sarkadi, my co-worshiper of the Machiavellian church. Collaborator, thinker, mate. To Reuth Mirsky, the 4th edge of ToM4AI for her mentorship and support, for her guidance. Last and by all means not least — Lion Schulz, forever my office mate. Lion and I shared both physical and mental office throughout our PhD. With Lion conversations about research flew smoothly from our heads, to the white board, to the computer and to papers. I'm forever in his debt for being my collaborator and my pal.

Lastly, my beloved family. My parents - Yair and Michal, who always supported me, even when my academic trajectory drifted. My parents in law - Shai and Michal, for supporting me and my family when I was traveling or absent due to deadlines. To my precious wife Yarden, the ultimate partner for this journey, my shelter, my support, my inspiration, who had to hear my endless ideas, pushed me and supported me in turbulent times. I couldn't have done this without you, my love. And my wonderful kids: Shahar, Yogev and Aluma - two of which were born during my PhD. My sunshine, my home, my inspirations.

Abstract

Deception is a constant in social life. From the house buyer who feigns indifference to lower the asking price, to the poker player who bluffs to inflate the stakes, agents routinely distort the signals they emit and discount the signals they receive. Far from being a special faculty, deception is what strategically sophisticated agents *do* when their interests conflict and information is asymmetric. This thesis investigates the cognitive and computational machinery that makes deception possible, asks what happens when that machinery misfires, and proposes a principled way to detect and deter deception without arms-racing into ever-deeper recursive reasoning.

The central cognitive ingredient is Theory of Mind — the ability to represent the beliefs, desires, and intentions of others, including their beliefs about one’s own mind. When an agent can model how an observer will interpret its actions, it gains the power to craft actions that install false beliefs in that observer. This capacity is recursive: a sufficiently sophisticated reasoner can model an opponent who is herself modeling the reasoner, and so on through nested levels of “I think that you think that I think.” This thesis shows that this recursion is both the engine of deception and the source of its most dangerous failure modes.

The first contribution establishes that deception and skepticism emerge from pure reward maximization when agents are endowed with recursive Theory of Mind, without any explicit programming of deceptive intent. In a signaling game where a buyer’s choices reveal information that a seller will use to set prices, a sufficiently sophisticated buyer discovers that distorting its choices is profitable. The ruse is not designed; it arises as a natural consequence of planning through the observer’s inference process. In turn, a more sophisticated seller, aware of this possibility, learns to discount the buyer’s signals in return — becoming strategically *skeptical*. Using information theory to measure these dynamics, we show precisely how much information the deceptive agent withholds and how much the skeptical agent discounts, providing a formal bridge between strategic behavior and the beliefs it is designed to manipulate.

The second contribution complicates the natural prescription implied by the first. If deeper recursive reasoning enables deception and its detection, should agents simply

maximize their depth of reasoning? The answer, it turns out, is no. When an agent applies a more sophisticated model than its opponent actually warrants, the same inferential machinery that correctly identifies a strategic opponent instead reads malicious intent into innocent behavior. The agent develops false beliefs of deliberate harm, reacts defensively to a naive partner, and suffers material losses that a well-calibrated agent would avoid. This failure mode — over-mentalizing — provides a formal computational account of paranoia: not as a deficit of social cognition, but as its pathological over-application. We identify the joint conditions that are both necessary and sufficient for this failure, show that the same mechanism can explain conspiratorial ideation and suspiciousness more broadly, and draw connections to clinical conditions including persecutory delusions and borderline personality disorder.

The third contribution asks whether there is a way to avoid being manipulated with finite and limited cognitive resources. The resolution lies in separating detection from modeling. Rather than trying to outthink a sophisticated deceiver — which is logically impossible for an agent with limited recursive capacity — a victim can instead monitor whether the deceiver’s behavior is consistent with any model the victim already possesses, and whether the rewards the victim is receiving match what those models would predict. When the answer to either question is no, the victim can conclude that it is facing an *unmodeled* opponent and switch to a defensive posture designed to make deception costly. We show that this simple augmentation substantially reduces the advantage a sophisticated deceiver holds over a limited victim, without requiring the victim to reason at a higher level.

Taken together, the three studies converge on a single organizing principle: the key variable governing whether Theory of Mind is *adaptive* or pathological is not the absolute depth of recursive reasoning, but its *calibration* relative to the opponent. When calibration is accurate, Theory of Mind is a precision instrument that enables strategic deception and warranted skepticism in equal measure. When it overshoots, it becomes a paranoia machine. When it undershoots, it leaves the agent open to exploitation. The computational framework developed here — and particularly the insight that agents can know that they do not know, and act accordingly — has implications for our understanding of psychiatric disorders rooted in social misattribution, for the design of AI systems that interact with strategic humans, and for the broader question of how intelligent agents can remain robust in a world populated by others who are trying to manipulate them.

Contents

Abstract	iv
Letter of Contribution	vii
1 Introduction	1
2 Paper I: A (Dis-)information Theory of Revealed and Unrevealed Preferences	22
3 Paper II: (Mal)adaptive Mentalizing in the Cognitive Hierarchy	40
4 Paper III: $\aleph$ -IPOMDP: Mitigating Deception in a Cognitive Hierarchy	60
5 Discussion and Conclusions	93
Bibliography	98

Letter of Contribution

This thesis is based on three published research papers. Below I describe my specific intellectual and technical contributions to each work. In all three studies I was the lead researcher responsible for the core scientific output.

Published as: Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind: Discoveries in Cognitive Science*, 7, 608–624. https://doi.org/10.1162/opmi_a_00097

Author contributions: N.A. and L.S. contributed equally as joint first authors. J.S.R. and P.D. contributed equally as joint last authors. N.A. developed the model and conducted all simulations. L.S. led the visualization. All authors contributed to conceptualization and writing.

Published as: Alon, N., Schulz, L., Bell, V., Moutoussis, M., Dayan, P., & Barnby, J. M. (2024). (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1), 159–177. <https://doi.org/10.5334/cpsy.117>

Author contributions: N.A. led data curation, formal analysis, investigation, methodology, software, and original draft preparation. L.S. contributed visualization. V.B. and M.M. contributed to writing (review & editing). P.D. and J.M.B. contributed to conceptualization, supervision, and writing.

Published as: Alon, N., Barnby, J. M., Sarkadi, S., Schulz, L., Rosenschein, J. S., & Dayan, P. (2026). $\aleph$ -IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection. *Journal of Artificial Intelligence Research*, 85, Article 14. <https://doi.org/10.1613/jair.1.19204>

Author contributions: N.A. is the corresponding author and led the formal framework, implementation, analysis, and writing. P.D. and J.S.R. led the

supervision. P.D. led the conceptualization. All co-authors contributed to formalization and writing.

Chapter 1

Introduction

“Never attempt to win by force what can be won by deception” — Niccolò Machiavelli, The Prince

Humans are social animals. We are programmed to live in a society [Banaji and Gelman, 2013]. We seek the help of others, we learn to communicate our needs [Bloom, 1993], we acquire the ability to cooperate with others to achieve greater goals together. We master skills and develop traits to navigate complex social interactions.

One of these underlying traits is Theory-of-Mind (ToM) [Premack and Woodruff, 1978, Apperly, 2012]. ToM is the cognitive trait that allows humans and other species (to a lesser extent) to fathom other agents’ latent cognitive processes. ToM serves multiple purposes in navigating complex interactions. First, it is a *generative model*. One can form beliefs about the thoughts of others: “I think that you think that...” [Kim et al., 2025, Baker et al., 2011]. As a result, ToM is used to make *inferences* about hidden intents [Baker et al., 2009] — if we observe a person walking towards the kitchen, we can infer that they are either hungry or thirsty as it is more likely than other alternatives. Lastly, ToM is used to plan optimal actions in social interactions [Ho et al., 2022, Hula et al., 2015]. The capacity to generate hypothesis, make inferences and compute optimal behavior, makes ToM a powerful trait.

ToM is hypothesized to emerge out of the need to manage (sometimes) conflicting interests of parties in large social complexes — like human society [Moll and Tomasello, 2007, Vygotsky and Cole, 1978]. This hypothesis helps us to understand why ToM is needed for *cooperative* tasks, like communication [Frank and Goodman, 2012], teaching [Shafra et al., 2014]. Effective communication between humans benefits from ToM [Mitchell, 2021]. In fact, human language acquisition and ToM co-evolve [Meltzoff, 1999], implying a tightly coupled link between our ability to express our thoughts and our capacity to predict how others will infer our words. ToM is used to cooperate

[Kleiman-Weiner et al., 2025], serving as the computational engine that allows agents to predict what others will do from limited observations, using the mental model of others to fill in the gap.

But there’s another, more sinister application of ToM in social lives. ToM is the apparatus of deception. Similarly to the way children learn to communicate their needs, they learn that sometimes bending the truth is beneficial [Mokkonen and Lindstedt, 2016, Crawford, 2003]. When we tell a lie, we are aware of two belief systems — ours and the victim’s. We know that what we’re saying isn’t true, and we assume that the victim believes us [Ward et al., 2023]. This is where ToM fosters deception — we can tell that we know something that others don’t (tracking belief systems) and we can use it to our benefit (planning). In fact, ToM and deceptive capabilities co-develop in kids [Chandler et al., 1989, Ding et al., 2015]. As the Italian Philosopher said — it is better to get what you desire by deception than by force.

However, as the proverb hints, victims may learn that they have been fooled and resent the ruse. This reaction not only degrades the deceiver’s ability to execute its subterfuge, it may trigger an arms race [Sarkadi, 2023], in which agents with higher levels of mentalization thrive at the expense of those with lower capacity.

Critically, the AI community has also explored and designed agents and systems that interact with one another. Similar traits to those examined in humans and other animals have been developed — from acquisition of communication protocols [Poslad, 2007] and interacting with new agents [Stone et al., 2010] to learning how to cooperate [Gupta et al., 2017]. The parallelism between the traits and features required to cooperate in multi-agent systems (MAS) suggests that the same underlying principles, and particularly ToM, can govern the interactions among and between artificial agents, and ultimately between artificial and natural agents.

This thesis seeks to explore the role of ToM in situations where deception is beneficial. It is composed of three chapters, each of which is a standalone paper, with the running thread of the role of ToM in games with asymmetric information. Overall, this work tells the story of how ToM endowed agents use this trait to manipulate others, how they utilize it to avoid manipulation with potentially dire outcome and how we can re-think ToM as a meta-system that is activated and deactivated in situations where it may hide the ToM-endowed agent.

Background

This work is positioned in the intersection of many domains (Fig. 1.1). In this section I will review the relevant literature with the aim of providing the reader with sufficient overview of the key concepts presented in this work.

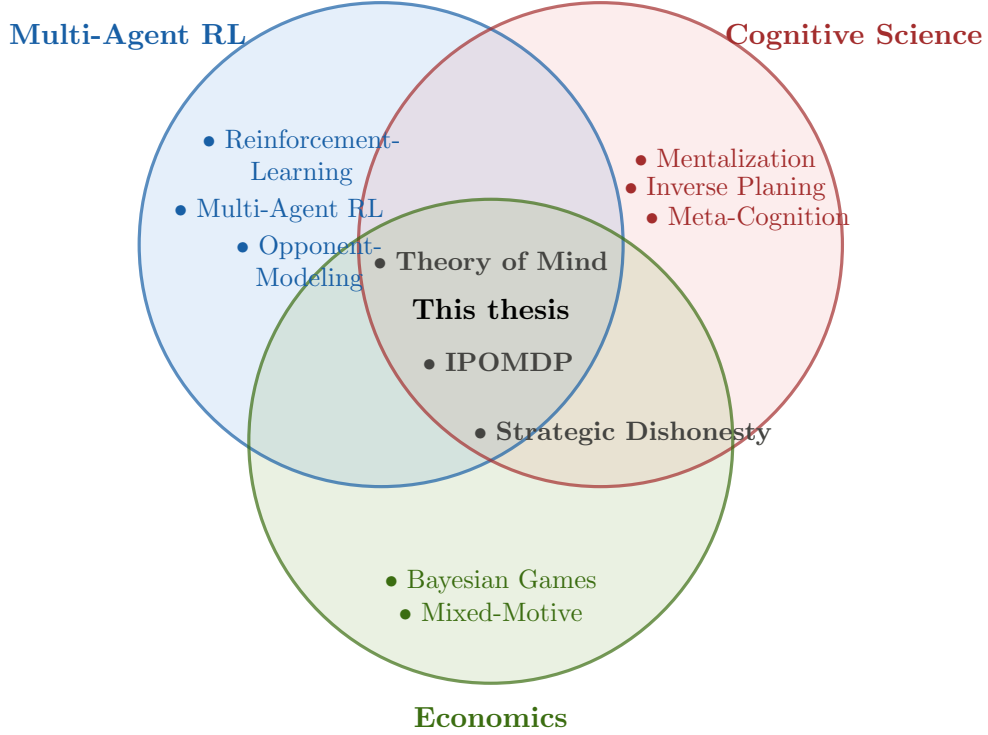


Figure 1.1: **Disciplinary positioning of the thesis.** The three circles represent the primary fields that this work draws on: Multi-Agent Reinforcement Learning, Cognitive Science, and Economics. The thesis sits at the three-way intersection, combining formal models of strategic interaction (*Economics*), recursive opponent modeling and Theory of Mind (*Cognitive Science*), and multi-agent planning and reward maximization (*Multi-Agent RL*).

The Link Between ToM and Reinforcement Learning

At the core of this thesis is the agent [Wooldridge and Jennings, 1994]. The agent interacts with an environment, which may host other agents as discussed next. The *state* of the environment describes the available information for the agent at a given time $t \geq 0$, and is denoted by $s^t \in S$. At each state, the agent takes an *action* $a^t \in A$. This action changes the state of the environment according to some (potentially unknown) dynamics: $s^{t+1} = f(s^t, a^t)$.

Agents have desires, things it wishes to accomplish. The agent’s goal is to find a sequence of actions that satisfies its desires. For example, if I have thirst (I have a desire for water), then I can walk to the kitchen and drink. In this work we assume that agents *learn* this sequence of actions using reinforcement learning (RL) [Sutton and Barto, 2018]. In RL, the agent receives a reward at each state, conditioned on the action it executed: $r^t = g(s^t, a^t)$, $r^t \in \mathbb{R}$, for some reward function $g()$. We assume that the reward corresponds to the agent’s *utility*. Since agents may differ in their utility (I like Brussels sprouts, but you hate them), we will address this difference when we present the multi-agent framework. For now, we use the terms reward and utility

interchangeably. A policy: $\pi : S \rightarrow A$ is a function that takes as input a state and outputs an action: $\pi(s^t) = a^t$. Each policy has a *value*, the expected discounted sum of rewards when the agent's actions are computed by π , starting from state $s^t \in S$:

$$V^\pi(s^t) = \sum_{k=0}^T E_{a^{t+k+1} \sim \pi} [\gamma^k r^{t+k+1} | s^t]$$

where $\gamma \in [0, 1]$ is known as the discount factor¹. Similarly, the action-value function (Q-value) defines the expected discounted sum of rewards, starting at state $s^t \in S$, taking action $a^t \in A$ and then following π for the remainder of the interaction:

$$Q^\pi(a^t, s^t) = \sum_{k=0}^T E_{a^{t+k+1} \sim \pi} [\gamma^k r^{t+k+1} | a^t, s^t] \quad (1.1)$$

The agent's goal is to learn (compute) the policy with the maximal value — known as the optimal policy:

$$\pi^* = \max_{\pi} \{V^\pi\}$$

The policy can be stochastic, a distribution over the actions for each state.

In multi-agent environments, the reward of one agent, r_i^t depends not only on its actions, a_i^t , but also on the actions of other agents, often denoted by a_{-i}^t . In this work we'll refer to the other agent as j , limiting the discussion to two agents:

$$V^{\pi_i}(s^t) = \sum_{k=0}^T E_{a_i^{t+k+1} \sim \pi_i} [E_{a_j^{t+k+1} \sim \pi_j} [\gamma^k r^{t+k+1}(a_i^{t+k+1}, a_j^{t+k+1})] | s^t] \quad (1.2)$$

where actions are performed simultaneously (like in RPS). If actions take place sequentially — like in chess — then the time index in Eq. 1.2 may be abused to define the within turn reward, but the action of j is now conditioned on i 's action:

$$V^{\pi_i}(s^t) = \sum_{k=0}^T E_{a_i^{t+k+1} \sim \pi_i} [E_{a_j^{t+k+1} \sim \pi_j | a_i^{t+k+1}} [\gamma^k r^{t+k+1}(a_i^{t+k+1}, a_j^{t+k+1})] | s^t] \quad (1.3)$$

In both cases the coupling implies that optimal policy computation depends on the agent's ability to model how other agents will act.

From Single-Agent to Multi-Agent RL There is an abundance of methods for modeling other agents in multi-agent systems [Albrecht and Stone, 2018]. In this work we focus on Bayesian inference based opponent modeling [Baker et al., 2009].

The *history*, h^t of the interaction is a vector containing the states and actions of

¹in this work we assume finite interactions

both agents: $h^t = (s^0, a_i^0, a_j^0, \dots, s^{t-1}, a_i^{t-1}, a_j^{t-1})$. Assuming that agents are *rational*, their action seek to maximize their utility, which is known privately (Eq. 1). Given the history, i can make inferences about j 's utility by *inverting* the behavior and using Bayesian inference:

$$p(u_j|h^t) \propto \prod_{k=0}^t P(a_j^k|u_j, s^k, h^k)p(u_j) \quad (1.4)$$

Where $P(a_j^k|u_j, s^k, h^k)$ is the likelihood of observing the observed sequence of actions, given the utility function, the states and i 's actions. $p(U)$ is the prior distribution over utilities and $p(u_j|h^t)$ is the posterior distribution over j 's utility. This is akin to the concepts of inverse reinforcement learning (IRL; [Ng et al., 2000]), naïve utility calculus [Jara-Ettinger et al., 2016], and [Ramirez and Geffner, 2009].

The underlying assumption of these methods is that the observed agent is *rational* (or Bayes-rational [Oaksford and Chater, 2007]). That is, the observee is trying to maximize its utility (reward), and the observer can test the hypothesis that a given utility function is the cause for the observed behavior.

Inverting the observed behavior implies that the observer has a *model* of the observee. That is, the observer can compute the optimal policy for each utility function.

From Inverse Planning to POMDP Notably, the utility of others is latent — only the observed agent *knows*² them while the observer *infers* them. Given that the observer is acting under uncertainty, we can model its decision making problem as Partially Observed MDP (POMDP; [Kaelbling et al., 1998]). In this scenario, the true state of the environment includes both the physical state s^t , as well as j 's utility. Had i known the utility, it could have solved Eq. 1.2 (assuming it can compute j 's optimal policy). Instead, i infers j 's utility from observations, forming a *belief* (distribution) over j 's utility: $b_i^t(u_j) = p(u_j|h^t)$ and use this information for optimal policy computation.

From POMDP to Interactive-POMDP The POMDP representation coupling physical state with opponent uncertainty is the core concept of Interactive-POMDP (IPOMDP; [Gmytrasiewicz and Doshi, 2005]). This framework models interactions with both environmental uncertainty (“vanilla” POMDP) and opponent uncertainty. In this thesis we focus on the latter, assuming that agents have full environmental knowledge, as in the framework presented by [Hoang and Low, 2013].

The *interactive state*, is_i represents the world from i 's perspective. As presented above, in the simple case, it includes the state s^t and j 's utility: $is_i^t = \langle s^t, u_j \rangle$ with the corresponding inference presented in Eq. 1.4.

²We assume that agents adhere to the scripture from the Temple of Apollo at Delphi.

Agents’ decisions are often governed by more than their utility. Limited computational resources, capped planning horizon, noisy behavior — all are factors that affect the agent’s ability to compute or execute the optimal policy. The collection of these traits — utility, bounded rationality, etc. — define the agent’s *persona* - ψ_i . Thus, the basic interactive state now includes j ’s persona: $is_i^t = \langle s^t, \psi_j \rangle$, retaining the inverse inference paradigm Eq. 1.4. The combination of the agent’s *persona* and *beliefs* $b_i^t(is_i^t) = p(s^t, \psi_j | h^t)$ defines the agent’s *type* [Harsanyi, 1968]: $\theta_i^t = \langle \psi_i, b_i^t(is_i^t) \rangle$.

Things become more complex when agents model other agents’ type rather than just the persona. Consider the case in which j is making an inference about i ’s type, this includes an inference about i ’s persona: $p_j(\psi_i | h^t)$ and also i ’s beliefs about j : $p_j(b_i^t(\psi_j | h^t))$.

$$b_j^t(is_j^t) = p(s^t, \psi_i, b_i^{t-1}(\psi_j) | h^t) \propto p(s^t | h^t) p(\psi_i \times b_i^{t-1}(\psi_j) | h^t) \quad (1.5)$$

Where the coupling $p(\psi_i \times b_i^{t-1}(\psi_j) | h^t)$ stems from the fact that the persona and belief are often hard to disentangle. Inferring the beliefs of another agent about the self creates a recursive structure — j infers how i perceives it. This is the recursion that is part of ToM (“I think that you think that I’m a friend”). Deeper recursed models follow the same belief update pattern as in Eq. 1.5, but with nested inference.

The number of recursed models defines the agent’s Depths of Mentalization (DoM; [Barnby et al., 2023]). At the core of the IPOMDP framework are the *subintentional* agents, also known as DOM(-1) agents.³ These agents lack explicit opponent model, and as such their environmental uncertainty (if exist) is limited to the physical state. Thus, these agents solve either a MDP or a POMDP. Higher levels of recursion correspond to increasing degrees of DoM, $k \in [0, \infty]$ — each of which represents other agents as having one DoM level beneath them.

K-level Reasoning and Some Alternatives In this work we make use of the IPOMDP framework, which follows the K-level model described above. The K-level reasoning (also known as K-level thinking) has been studied extensively in economics [Benndorf et al., 2017, Arad and Rubinstein, 2010, Ho and Su, 2013, Yoshida et al., 2008, Nagel, 1995], where it is used to describe how bounded rational agents (humans) make decisions when facing other agents with bounded rationality. Other models, like the Cognitive Hierarchy (CH) [Camerer et al., 2004] model allow the agents to form a (truncated) distribution over the various k values — nonetheless, all models assume that a DoM(k) agent cannot reason about agents with higher DoM [Pacuit and Roy, 2017].

³The author is aware of the ongoing debate whether DoM begins at 0 or -1.

Solving IPOMDP Computing the optimal solution for POMDP is PSPACE-Complete [Bernstein et al., 2002, Tsitsiklis and Athans, 2003]. Since IPOMDPs are recursed POMDPs, IPOMDP is PSPACE-hard (for finite time horizons). This implies that approximated solutions are needed to solve IPOMDPs in a reasonable time. One such method is the Interactive-POMCP (IPOMDP; [Hula et al., 2015]). This sample based algorithm augments the POMCP algorithm [Silver and Veness, 2010] for IPOMDP. Briefly, the algorithm expands the game tree by sampling particles [Doshi and Gmytrasiewicz, 2009] of interactive states and evaluating the action-value (Q-values; $Q(a^t, s^t) = r(a^t, s^t) + V^\pi(s^{t+1})$) of the action. During computation the algorithm traverses the tree, favoring higher reward paths. after an action is taken, the tree is pruned, with the new root corresponding to the current history. The particles associated with this history are used to estimate the belief and the policy computation continues.

The Link Between ToM and Game Theory

Modeling strategic interactions between agents has been extensively explored in Game Theory [Maschler et al., 2020]. In this thesis we focus on a particular type of interaction: games with incomplete information, also known as Bayesian games [Harsanyi, 1968]. The goal of this section is to draw parallels between Bayesian games and the IPOMDP representation, as well as to present the reader with the concept of Mixed-Motive games.

Bayesian games [Zamir, 2020] are strategic interactions in which the acting agents have limited (partial) information about the game. Agents (players) may be misinformed about the payoff of the game, or about the strategies of other agents. The various forms of environmental uncertainty in Bayesian games correspond to the same environmental uncertainties represented in the IPOMDP framework.

Overcoming the interaction ambiguity gives rise to a *hierarchy of beliefs*, which corresponds to recursive opponent modeling — one of the features of ToM. We provide a minimal presentation of Bayesian games and hierarchy of beliefs and refer the reader to [Zamir, 2020] for a full exposition of the topic.

Consider the following game (Fig. 1.2) (taken from [Zamir, 2020]). Two agents, i and j play *one* of the following games:

The agents do not know which game they play, this information is concealed (selected by nature). Had the agents known which game they play (the true state of the world), they could have computed the optimal solution (Mixed-Nash Equilibrium; MNE). Given that they do not know which payout matrix is sampled, they form beliefs about the game. Denote by p the prior probability the row agent assigns to the event that $G1$ is the game: $p = P(s = G1)$. Similarly, denote by q the probability assigned to the same event by the column agent.

		Column agent	
		L	R
Row agent	T	(0,1)	(1,0)
	B	(1,0)	(0,1)

(a) $s = G_1$

		Column agent	
		L	R
Row agent	T	(1,0)	(0,1)
	B	(0,1)	(1,0)

(b) $s = G_2$

Figure 1.2: Payoff matrices for games G_1 and G_2 . Entries denote the agents' payoffs (R,C).

Row belief	Column agent's action	
	L	R
$p < 0.5$	T	B
$p > 0.5$	B	T

(a) Row optimal action

Row agent	Column belief	
	$q < 0.5$	$q > 0.5$
T	R	L
B	L	R

(b) Column optimal action

Figure 1.3: Optimal action matrices for games G_1 and G_2 and beliefs p, q . Entries denote the agent's optimal action given their first-order beliefs.

A central solution concept for Bayesian games is the *Bayes-Nash Equilibrium* (BNE; [Harsanyi, 1968]). A BNE is a strategy profile in which each agent's strategy is a best response to the strategies of all other agents, given its beliefs about the unknown state of the world and given that those beliefs are derived from a common prior via Bayes' rule. Formally, a strategy σ_i for agent i is part of a BNE if, for every type of agent i :

$$\sigma_i^* \in \arg \max_{\sigma_i} \mathbb{E}_{s \sim p(\cdot | \theta_i)} [u_i(\sigma_i, \sigma_{-i}^*(s), s)] \quad (1.6)$$

where θ_i is i 's type (encoding its private information and beliefs), s is the true state of the world, and the expectation is taken over i 's posterior beliefs about s given θ_i . The BNE thus extends the Nash Equilibrium to settings of incomplete information by requiring rationality to hold *in expectation* over the unknown state, rather than at each realized state. In the game of Table 1.2, neither agent knows which matrix is being played, so the equilibrium strategies must be optimal on average over the agent's posterior beliefs — which, as the best-response tables in Figure 1.3 show, depend critically on the beliefs p and q . The IPOMDP framework can be understood as extending the BNE to sequential, partially observable settings: agents maintain recursive beliefs about each other's types and plan policies that are optimal in the BNE sense at every decision point along the interaction [Gmytrasiewicz and Doshi, 2005].

The best response is the action that will maximize the agent's utility (reward) given the actions of the other agent. To compute it, we use the same blueprints presented in the previous section. The row agent's optimal action depends on its beliefs (p) and the column agent's action in each scenario:

Row’s utility depends on the actions of Column, which depends on Column’s beliefs about the game played (q). Thus, Row can form second-order beliefs — a belief about Column’s beliefs. This structure of nested beliefs is comparable to the recursive modeling in IPOMDP. This parity, which was acknowledged by [Gmytrasiewicz and Doshi \[2005\]](#), allows us to use both framework to model interactions with partial information.

From Bayesian Games to Signaling Games Signaling games [[Sobel, 2020](#)] is a family of Bayesian games with sequential interactions. These games retain the asymmetric information structure and the Bayesian rationality paradigms of Bayesian games, while allowing for strategic behavior to emerge. Particularly, signaling games model how agents hide their type to improve the outcome — a model of deception.

We use [Spence \[1978\]](#)’s seminal work on signaling in the job market for the exposition. In [[Spence, 1978](#)] a firm is hiring a worker for a job. The worker may be hard working (H) or lazy (L). If the firm hires a lazy employee, it is harmed, while the worker is happy, if the firm hires a hard working worker, it benefits, but the worker gains less reward than its lazy counterpart — as it is putting effort into its work.

The true character of the worker is known only to them (akin to *persona* in IPOMDP), and the company can infer this feature from the worker’s level of education. A worker can earn a degree, which *signals* its quality. Hard working agents suffer less for gaining the title while the lazy worker suffers more. Hence, the firm infers the persona ψ_i (lazy or hard working) from the workers decision to pursue an academic degree a_i :

$$p(\psi_i|a_i) \propto P(a_i|\psi_i)p(\psi_i)$$

This inference is similar to the Bayesian inverse planning inference and corresponds to DOM(0) inference presented above. Thus, we summarize this part with the observation that these models describe the same phenomena — strategic self-centered agents, with private information, that decide which signal to share with an observer to their benefit.

From Bayesian Games to Hypergame Hypergame theory, introduced by [Bennett \[1980\]](#), extends classical game theory to settings where players hold diverging or incomplete views of the game being played — precisely the structural conditions under which deception thrives.

Unlike standard Bayesian games, where agents share a common prior over payoff matrices, a hypergame allows each player to perceive a different version of the conflict, capturing misperception and intentional deception within a unified formal framework.

The theory has been applied to model deception [[Kovach et al., 2015](#)] in several domains: in information warfare, where [Kopp \[2002\]](#) showed that denial, degradation,

and corruption of information channels map directly onto hypergame structures; in cybersecurity, where attacker-defender models exploit the asymmetry between what each side believes the game to be; and in conflict analysis, where the hypergame perception model was used to reconstruct historical episodes of strategic deception such as the Cuban Missile Crisis and the Fall of France.

The central mechanism in all these cases is the same — a deceiver with superior information exploits the victim’s misconception of the game to achieve outcomes that would be unavailable if both parties shared the same game model. For this thesis, the key takeaway is that hypergames provide an independent and complementary formalism to the IPOMDP for modeling deception: where the IPOMDP frames deception as belief manipulation through nested recursive modeling, hypergame theory frames it as the exploitation of divergent game perception, and the two converge on the same structural insight — deception requires and feeds on an asymmetry in the cognitive models agents hold of their shared interaction.

Mixed-Motive Games A core concept in this thesis is Mixed-Motive Games. Mixed-motive [Gallo Jr and McClintock, 1965] games are ones in which agents are conflicted between their individual gains and the necessity to satisfy others’ selfish needs. These games include many social conflicts from negotiation [de Weerd et al., 2017, Hodayun, 2018] to investments [Hula et al., 2018]. In particular, this thesis focuses on uncertainty about other agents’ types. That is, imagine a seller selling multiple items (like large retailers). Ideally, the price of the items should match the willingness of the customers to pay for the item — a Coke addict would pay more for the drink than a Pepsi fan. However, the preferences of the customers are known only to them, limiting the retailers’ ability to adapt the price. Hence, the *type* of each customer describes its Coke-Pepsi preference, and the vendor is trying to make inferences about this unknown preference (type) in a Bayesian manner [Ray et al., 2008]. By inverting the decisions of the customers (which beverage they add to their cart, the trader can update its *beliefs* about their taste and set the price. If the customers know this — that they are being observed [Miura and Zilberstein, 2024], and they know that this information could be used against them (as the price of their favorite refreshment spikes), they are incentivized to avoid disclosing this information.

The Link Between ToM and Deception

The link between ToM and deception is based on empirical findings and theory from disparate branches of science. In a nutshell, the combination of ToM and Mixed-Motive Bayesian games provides us with the setup for the emergence of deception — conflicting interests, falsifiable beliefs and recursive opponent modeling [Ettinger and Jehiel, 2010,

Aitchison et al., 2021].

Previous work focused on deception with low levels of DoM. For example, observer-aware agents (corresponding to DoM(0)) learn to hide their intents or manipulate the perception of others (planting false beliefs). This phenomenon was explored in developmental psychology [Sodian et al., 1991] where children learn that they can utilize others’ uncertainty about the whereabouts of an item to their benefit. In MAS, it has been shown that agents avoid goal detection [Masters and Sardina, 2017] by hacking the observer’s beliefs and increasing the observer’s uncertainty about their target. The AI community became very interested in modeling deceptive behavior in Bayesian games in the context of Cybersecurity [Huang and Zhu, 2018]. Deceptive behavior emerges as the result of solving games like Poker [Brown and Sandholm, 2018, Southey et al., 2012]. Our work extends this blueprint by quantifying and measuring the effect of deceptive ploys on other agents’ beliefs — providing a needed bridge between observed outcome (ruse, bluff) and the causal model enabling the success of the gambit — false beliefs and corresponding behavior. Other papers modeled deception with higher levels of DoM [Hula et al., 2018], from which we draw inspiration in Chapter 3.

Coping with deception is where ToM becomes handy. While deception can be “accidental” or unintended, telling that one is interacting with a deceptive opponent is harder [Sip et al., 2008]. First, if agents are uninformed that they are matched with deceptive others, they need to fathom that this might be possible. Second, on a higher level, the agent needs to detect that this situation is one where deception is beneficial — a preemptive step to detecting deception. This thesis shows how ToM is used for both tasks simultaneously — by simulating the minds of others in the game, ToM endowed agents use their nested models of other to detect when they are being deceived (or at least *think* that they are being deceived), and how they adapt their behavior to mitigate the risk of harm.

One of the limitations of ToM agents is their inability to reason about agents with higher cognitive ability (higher DoM). As such, the ploys of higher DoM agents rely on this limitation. One possible way to overcome this limitation, is using a mechanism that can either *adapt* the agent’s DoM [Yu et al., 2022]. However, these models are based on best-response dynamics, which makes the adapted agent susceptible to manipulation, thus not mitigating the risk of being deceived. Sometimes the deceiver’s behavior fails to adhere to *expected* behavior [Hula et al., 2018]. This approach is related to open-set recognition (OSR) [Vaze et al., 2022]. In OSR the agent is given a set of *recognized* categories and is tasked with filtering new categories during test time. In the context of this work, this is analogous to a ToM agent that detects unexplained actions during interaction. That is, ToM is used as a generative model, and none of the models (types) produces the observed behavior.

The Logic and Philosophy of Deception Defining deception precisely is harder than it first appears, and the difficulty is illuminating. The traditional philosophical definition of a lie requires four conditions: a statement is made, the speaker believes the statement to be false, the statement is directed at an addressee, and the speaker intends the addressee to believe the statement to be true [Mahon, 2018, Isenberg, 1973]. Yet this definition is both too narrow and too broad.

It is too narrow because it restricts deception to explicit false statements, excluding the many forms of manipulation that operate through true but misleading utterances — what philosophers call *paltering* or *false implicature* [Mahon, 2018]. It is too broad because the falsity of the statement is not what makes a lie a lie. What matters is the speaker’s *belief* that the statement is false, not whether the statement actually is [Mahon, 2018, Ward et al., 2023]. A speaker who sincerely asserts something false is not lying. A speaker who asserts something true while believing it to be false is. This asymmetry — that the moral and epistemic weight of deception falls on the manipulation of the *victim’s* beliefs, not on the truth value of the signal itself — is precisely the structure that theory of mind is needed to exploit. Deceiving requires tracking not just what one *knows*, but what the victim currently *believes* and what they will come to believe given an observed action.

The philosophical literature draws a further distinction between *lying* and *deceiving* that is especially relevant here. Lying requires a statement, deceiving does not [Mahon, 2018]. One can deceive through action, omission, tone, timing, or selective disclosure — any signal that causes a false belief in the target. This broader category is what the present thesis models: agents that deceive not through false utterances but through strategically chosen *actions* that manipulate the observer’s Bayesian inference about their type. The buyer in *Paper I* never asserts anything false; it simply selects a path through the maze that causes the seller’s inverse reinforcement learning to converge on the wrong posterior. The DoM(1) sender in *Paper II* and *Paper III* never claims to be random; it merely behaves in a way that makes randomness the most likely explanation under the receiver’s model. These are paradigm cases of deception without lying — what the philosophical tradition calls *misleading* [Mahon, 2018, Saul, 2012] — and they are formally tractable precisely because they operate through the manipulation of a well-defined probabilistic belief update.

A third distinction from the philosophical literature concerns the relationship between deception and *intention*. Mahon [2018] following Chisholm and Feehan [1977] requires that a deceiver not merely cause a false belief but do so *intentionally* and with an *ulterior motive* — the false belief must tip the victim’s behavior in a direction that benefits the deceiver. This tripartite structure (false belief, intentional causation, behavioral consequence) maps directly onto the axiomatic definition of deception

developed in *Paper III*: Incentive Compatibility captures the ulterior motive, Epistemic Manipulation captures the installation of false beliefs, and Behavioral Causality captures the requirement that the false belief produce regret-inducing actions. The fourth axiom, Cognitive Dominance, adds the distinctly computational constraint that is absent from most philosophical accounts: deception of this kind requires that the deceiver possess a strictly higher recursive model of the victim than the victim possesses of the deceiver. Philosophy tells us what deception is; the cognitive hierarchy tells us what computational resources are required to execute it.

Measuring Deception In this work we define *deliberate* deception as an adversarial attack on the victim. Particularly, we are interested in the manipulation of beliefs, from which the agents compute their policy. To quantify how the inference is affected by the adverse ploy, we follow Kopp et al. [2018] and measure deception using methods from Information Theory (IT) [MacKay, 2003]. Specifically this thesis presents and analysis of two types of deception, introduced by Kopp et al. [2018]: *degradation* and *corruption*. Degradation is defined as when a player “conceals or hides information in noise, or other background messages, to introduce uncertainty or a false perception in a competing player’s belief”. Corruption is defined when a player “produces a false belief by replacing a real message with a similar, but false message, contrived to be very difficult to distinguish from a real message”. In Chapter 1, we measure these quantities using Mutual Information (MI) and the Kullback–Leibler Divergence (KLD). MI is a metric defined between two random variables, X, Y :

$$I(X; Y) = H(X) - H(X|Y)$$

where $H(X)$ is the *entropy* of a random variable X and $H(X|Y)$ is the conditional entropy of X given Y . MI measures how much uncertainty about X can be reduced by observing Y . KLD is a function defined between two distributions, P, Q , (defined over the same support X):

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{Q(x)}{P(x)}$$

KLD measures how much an approximation distribution Q differs from the true distribution P .

In the context of deception, this quantity measures the discrepancy caused by modeling an observed behavior using one model (observer model) Q , when the behavior is actually generated by another model P . That is, we use this size to measure how “wrong” the inferring agent’s model of the observee is, relative to the observee’s true model. Overall, we make use of IT metrics to assess the magnitude of deceptive behavior on the perception (beliefs) of the victim.

The Link Between ToM and other Cognitive Traits

ToM does not operate in isolation. It is tightly coupled with a set of higher-order cognitive capacities that regulate *when* and *how* it is deployed. Chief among these is *metacognition* — the ability to monitor and evaluate the quality of one’s own cognitive processes [Fleming and Dolan, 2012, Nelson and Narens, 1990].

The relationship between ToM and metacognition is bidirectional. On the one hand, ToM presupposes a form of metacognitive awareness: to model another agent’s belief state, an agent must first have some representation of belief as a mental category — a capacity that requires reflecting on mental states rather than merely holding them [Frith, 2012]. On the other hand, metacognition is needed to regulate ToM itself. An agent that can assess the reliability of its own opponent model — asking “am I confident that I understand what my opponent is doing?” — can decide when to trust its ToM-derived inferences and when to suspend them [Schulz et al., 2023b]. This regulatory role is precisely what the $\aleph$ -mechanism in *Paper III* formalizes: it is a computational implementation of metacognitive monitoring applied to opponent modeling, detecting when the agent’s world model is failing and triggering a switch to a model-free defensive posture.

Empirically, metacognition and ToM are dissociable but correlated. Individual differences in metacognitive sensitivity predict social inference accuracy across tasks [Buergi et al., 2024]. Crucially, metacognitive failures — the inability to accurately assess one’s own epistemic state — have been linked to the same psychiatric populations that show ToM abnormalities, including schizophrenia [Lysaker and Dimaggio, 2014] and borderline personality disorder [Bateman and Fonagy, 2010]. This convergence suggests that the pathological social inference patterns described in this thesis may reflect not just a miscalibrated ToM, but a failure of the metacognitive layer that would ordinarily correct it.

Executive function is a further cognitive trait intertwined with ToM. Working memory, cognitive flexibility, and inhibitory control all support the recursive computations that higher DoM reasoning demands [Carlson and Moses, 2001, Devine and Hughes, 2016]. Developmental studies confirm that ToM and executive function co-mature in childhood, and that executive function limitations constrain the depth of recursive reasoning children can sustain [Perner and Lang, 1999]. In adult populations, individual differences in working memory capacity predict performance in exactly the kind of iterated strategic games used in this thesis [Ghosh and Gilboa, 2014]. These findings motivate a resource-rational perspective on DoM [Bhui et al., 2021]: agents do not simply choose their level of recursive reasoning; they are constrained by the cognitive resources available to support it. The $\aleph$ -mechanism can be understood from this perspective as a

heuristic that conserves those resources — substituting pattern matching over observed behavior for the computationally expensive process of full recursive inference.

Computational psychiatry The findings of this thesis contribute to a growing body of work using formal computational models to understand psychiatric conditions rooted in aberrant social cognition [Montague et al., 2012, Huys et al., 2016]. Abnormal ToM has been implicated across a wide spectrum of disorders, and the framework developed here offers mechanistic accounts for several of them.

Paranoia and persecutory delusions are the most direct application of *Paper II*. Over-mentalizing — miscalibrating the DoM levels of the self and other — produces the signature features of paranoia: false attribution of malicious intent to innocuous behavior, rigid updating, and behavioral consequences that confirm the false belief by preventing disconfirmatory evidence from accumulating [Freeman, 2016b, Garety et al., 2001]. The model predicts that paranoid individuals will show systematically elevated effective DoM when playing against simple partners in the Iterated Ultimatum Game [Xiang et al., 2013], a prediction that is testable with established paranoia scales [Freeman et al., 2005].

Borderline personality disorder (BPD) is characterized by hyper-mentalizing — making excessively convoluted inferences about others’ social cues [Fonagy and Bateman, 2006, Sharp et al., 2011]. This maps directly onto the over-mentalizing failure mode of *Paper II*: BPD individuals are theorized to attribute an excessively high level of intentionality to sparse social data, producing the unstable interpersonal representations that are the clinical hallmark of the disorder [King-Casas et al., 2008]. The present framework formalizes this as a DoM miscalibration rather than a global ToM deficit, consistent with empirical evidence that BPD individuals perform normally on standard false-belief tasks [Sharp et al., 2011].

Autism spectrum conditions (ASC) present the mirror-image failure mode: under-mentalizing, or insufficient ToM depth to model the intentions of typically-developing interaction partners [Baron-Cohen et al., 1985, Frith and Happé, 1994]. The $\aleph$ -mechanism of *Paper III* is relevant here: individuals who cannot fully model a more sophisticated partner may nonetheless be able to detect that their expectations are being violated, and respond protectively.

Schizophrenia involves a complex pattern of ToM abnormalities, including both under-mentalizing on standard tasks and over-attribution of agency in more naturalistic settings [Penn et al., 2008, Barch and Ceaser, 2012]. The dual-route model implied by *Paper II* and *Paper III* — paranoia arising from either excessive recursion or over-sensitive anomaly detection — may help reconcile these apparently contradictory findings, suggesting that different patient subgroups or symptom profiles correspond to

different points of failure in the social inference pipeline.

Taken together, these applications position the thesis at the intersection of formal AI methods and clinical cognitive science, with the potential to generate testable predictions across diagnostic boundaries. Future experimental work should deploy the IUG and buyer–seller paradigms in clinical populations, using computational model fitting to estimate individual DoM levels and anomaly-detection sensitivities, and relate these parameters to established symptom measures.

Following the general background and positioning of this work, I now present a summary of each paper, discussing its contribution to our understanding of the various roles ToM play in Mixed-Motive Bayesian games.

Paper I: The Emergence of Deception and Skepticism [Alon et al., 2023]

The first study asks a fundamental question: can deception and skepticism *emerge* from pure reward-maximization in a Mixed-Motive, Bayesian game without being hand-coded into any utility function? The answer is yes — and the mechanism is transparent enough to be analyzed with information theory.

This paper contributes two main results. First, ToM is a *disinformation machine*. In signaling games [Sobel, 2020, Sasahara and Sandberg, 2022], ToM endowed agents us this trait to pick actions (signals) that distorts the belief of the observer, and they do so to gain a better outcome. The second outcome is a detailed Game Information analysis of this distortion. We measure not only how much belief corruption [Kopp et al., 2018] takes place, but also quantify the discrepancy between presumed model and reality. In addition we quantify *skepticism* as the lack of belief update in Bayesian ToM (B-ToM) [Baker et al., 2011].

Paper II: When Recursion Misfires — the Link to Paranoia [Alon et al., 2025]

The second chapter answers the question posed by the Cognitive Arms race hypothesis: is the remedy for deception is to have higher DoM than anyone else? The K-level model is only beneficial when an agent’s depth of mentalizing (DoM) is calibrated to its opponent’s. When levels mismatch, *over-mentalization* takes place — actions are interpreted through the lenses of sophistication with corresponding (potentially detrimental) response. This research paves the way to viewing some Psychiatric disorders as stemming from maladapted ToM.

The study draws a formal link between over-recursive social cognition and clinical paranoia, persecutory delusions, and borderline personality disorder — conditions characterized by the attribution of excessive, malicious intentionality to innocuous behavior. It provides a mechanistic account rooted in reward maximization and Bayesian inference, rather than deficits of ToM.

Paper III: Counter-Deception Without Deeper Recursion [Alon et al., 2026]

The first two studies establish the problem: deception is an emergent feature of the cognitive hierarchy. Lower DoM agents will be manipulated by those with higher DoM levels without the ability to infer that this is the case. The third study offers a solution that does not require victims to rise higher in the hierarchy — an impossibility given the logical constraints of self-referential modeling [Pacuit and Roy, 2017, Gmytrasiewicz and Doshi, 2005].

I present an augmentation of the IPOMDP [Gmytrasiewicz and Doshi, 2005] framework to account for open-set world [Vaze et al., 2022]. This framework, the $\aleph$ -IPOMDP, allows agents to detect when they are matched with unmodeled others and avoid deception by adopting model-free retaliatory behavior. Tested in both a mixed-motive and a zero-sum game, the $\aleph$ -IPOMDP substantially reduces the reward gap between the deceiver and the victim.

This paper contributes to our view of ToM as both a generative and inference mechanism that can (and should) be used when properly calibrated and avoided when not.

A Unified Overview: Calibration as the Key Variable

Figure 1.4 summarizes the main layout of this work. Across all three studies, the critical variable is not *depth* of mentalizing per se, but its *calibration* relative to the actual opponent. Notably, if one agent is *calibrated* the other is *miscalibrated* by definition — here we use calibration from the self perspective.

When the self DoM is well calibrated with the opponent’s DoM, ToM is a precision instrument: agents can model each other’s inference processes, produce strategic deceptions, and develop warranted skepticism — all from pure reward-maximization, with information theory providing the measurement language (*Paper I*). When calibration overshoots, ToM becomes a paranoia machine: the same Bayesian update that in the well-calibrated case correctly identifies a deceptive sender instead attributes malicious

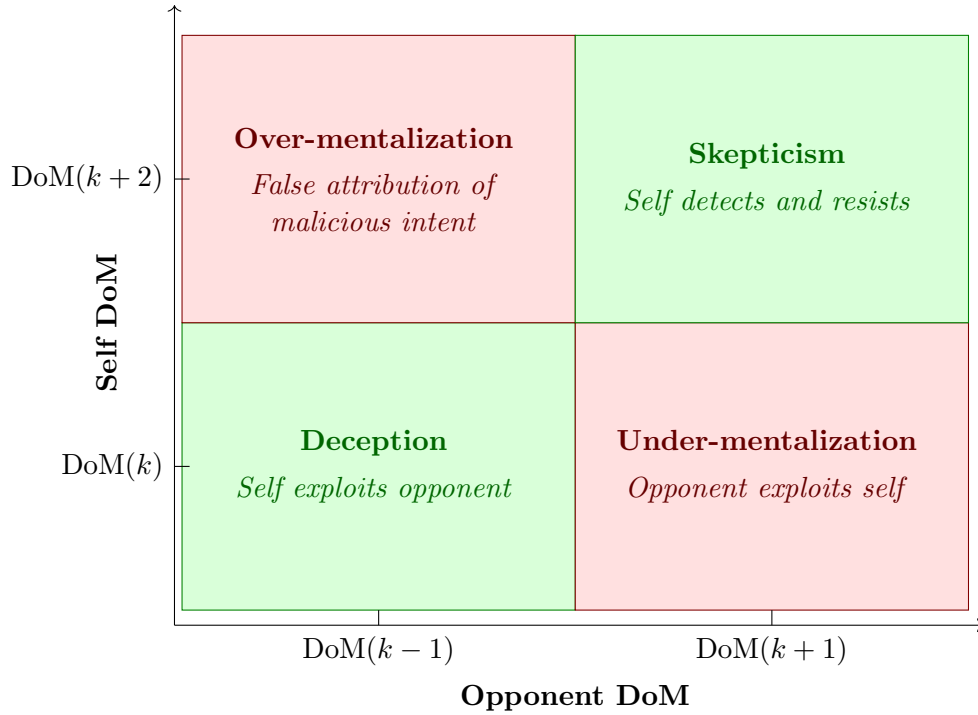


Figure 1.4: **ToM Miscalibration Matrix:** The matrix cross-tabulates self and opponent Depth of Mentalizing (DoM). Green cells (main diagonal) correspond to adaptive outcomes: a DoM(k) self facing a DoM($k-1$) opponent gives rise to *deception*, while a DoM($k+2$) self facing a DoM($k+1$) opponent gives rise to *skepticism* [Alon et al., 2023]. Red cells (off-diagonal) correspond to maladaptive outcomes: a DoM($k+2$) self facing a DoM($k-1$) opponent leads to *over-mentalization* and paranoid-like false attributions [Alon et al., 2025], while a DoM(k) self facing a DoM($k+1$) opponent leads to *under-mentalization* and unchecked exploitation [Alon et al., 2026].

intent to innocent behavior, incurring both material loss and false negative inference about the world (*Paper II*). When calibration is insufficient — when an agent lacks the recursion depth to model its opponent — deception succeeds unchecked. The $\aleph$ -mechanism offers an escape hatch: not by ascending the hierarchy, but by monitoring the *symptoms* of being deceived (atypical behavior and lower-than-expected reward) and responding with a, although potentially costly, credible deterrent (*Paper III*).

Broader Implications

This work is rooted in multiple disciplines and as such has applications to several domains.

Computational psychiatry. The findings of *Paper II* and *Paper III* converge on a transdiagnostic account of paranoia and other psychiatric disorders like delusional disorder and BPD. Whether arising from miscalibrated recursive modeling or from over-sensitive anomaly detection, the clinical signature is the same: false attribution of malicious intentionality to ambiguous or innocent behavior. This provides two distinct computational mechanisms for the same phenotype, consistent with the view that paranoia is not a unitary disorder but a family of inferential biases [Barnby et al., 2023]. Future experimental work can dissociate these mechanisms by independently manipulating opponent sophistication and reward noise in the IUG.

AI safety and alignment. The emergence of deception from pure reward-maximization in *Paper I* carries a direct warning for AI systems that interact with humans or other agents: a sufficiently recursive model-based agent *will* discover deceptive strategies if they are instrumentally useful, without deception being explicitly programmed. With the recent surge in LLM deceptive behavior, [Hagendorff, 2024] and the ongoing exploration of the emergence of ToM in LLMs [Kosinski, 2023] this becomes a pressing issue. The $\aleph$ -IPOMDP of *Paper III* offers a blueprint for detecting such behavior: an external monitoring layer that evaluates whether an AI agent’s outputs are consistent with its expected policy [Fleming and Dolan, 2012] and whether the human’s accumulated utility is within expected bounds.

Multi-agent RL. The thesis contributes a formal, axiomatic definition of deception in multi-agent RL (*Paper III*), grounded in four necessary conditions: incentive compatibility, epistemic manipulation, behavioral causality, and cognitive dominance. This formalization is task-agnostic and extends naturally beyond the games presented to any Bayesian repeated game.

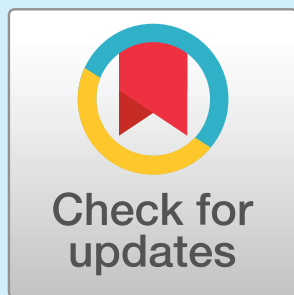
The thesis is accompanied by open-source code for all simulations.

Chapter 2

Paper I: A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind

Published as: Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind: Discoveries in Cognitive Science*, 7, 608–624. https://doi.org/10.1162/opmi_a_00097

Author contributions: N.A. and L.S. contributed equally as joint first authors. J.S.R. and P.D. contributed equally as joint last authors. N.A. and L.S. developed the model and conducted all simulations. All authors contributed to conceptualisation and writing.



Citation: Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind: Discoveries in Cognitive Science*, 7, 608–624. https://doi.org/10.1162/opmi_a_00097

DOI:
https://doi.org/10.1162/opmi_a_00097

Received: 21 April 2023
Accepted: 19 July 2023

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Lion Schulz
lion.schulz@tue.mpg.de

Copyright: © 2023
Massachusetts Institute of Technology
Published under a Creative Commons
Attribution 4.0 International
(CC BY 4.0) license



The MIT Press

REPORT

A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind

Nitay Alon^{1,2*} , Lion Schulz^{2*} , Jeffrey S. Rosenschein^{1†} , and Peter Dayan^{1,3†} 

¹Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel

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

³Department of Computer Science, University of Tübingen, Tübingen, Germany

*Denotes joint first-authorship.

†Denotes joint last-authorship.

Keywords: theory of mind, deception, communication, skepticism, disinformation, multi-agent-RL, IPOMDP

ABSTRACT

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

INTRODUCTION

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

Theory of Mind as Inverse Reinforcement Learning

In machine learning, inferring an agent's preferences, or utility function, is part and parcel of models that can be broadly summarized as 'inverse reinforcement learning' (IRL; Ng & Russell, 2000). These models observe an agent's actions and try to deduce the agent's preferences from these observations. The core insight of these models is that agents will only perform actions that

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

In cognitive science, inverse reinforcement learning has been used for models of the powerful inferences humans draw about one another (Baker et al., 2017; Jara-Ettinger, 2019). Chiefly among them, the ‘Naïve Utility Calculus’ (Jara-Ettinger et al., 2016) proposes that humans reason in ways similar to Bayesian inference. This type of model has successfully explained how we reason about one another’s kindness, knowledge, effort allocation, and skills (Berke & Jara-Ettinger, 2021; Xiang et al., 2023). Interestingly, this inferential ability arises early in development, with children as young as 5 years old being able to make intricate inferences (Jara-Ettinger et al., 2015). Across fields, this ability to reason about others’ beliefs, desires and intentions has been referred to as ‘theory of mind’ (Premack & Woodruff, 1978), a trait hypothesized to be foundational to the human ability to carry out complex social interactions (Camerer et al., 2015; De Martino et al., 2013; Ho et al., 2022; Hula et al., 2015; Ray et al., 2008; Rusch et al., 2020; Steixner-Kumar et al., 2022; Xiang et al., 2012).

Extending Inverse Reinforcement Learning Through Recursion

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

Such recursive reasoning about other agents extends theory of mind and has been used to explain a number of different phenomena in human interaction, from how we teach to how we lie and wage war (Crawford, 2003; Ho et al., 2022; Oey et al., 2023). Broadly, formal models of this recursivity (Doshi et al., 2020) tend to extend the simple inference in inverse reinforcement learning, and ‘plan through’ this inference model. A particularly flexible instantiation of this is the multitagent reinforcement learning framework involving Interactive Partially Observable Markov Decision Processes (IPOMDP; Gmytrasiewicz & Doshi, 2005). In an IPOMDP characterization, agents reason not only about uncertainty in the environment (as in a regular POMDP), but also about other agents’ beliefs, desires, and intentions. Furthermore, agents do so in a recursive manner, at different levels of a ‘cognitive hierarchy’ (Camerer et al., 2004). That is, agents of different sophistication model agents that are less sophisticated than themselves (‘I believe what you believe what I believe’ and so on). On average, humans are hypothesized to reason to a depth of 1.6 levels (Camerer et al., 2004).

Information Theory in Multi-Agent Interactions

In an IPOMDP, one agent’s actions are interpreted by other agents as signals about the hidden or latent characteristics that would be the target of IRL. Information theory (IT) is a particularly helpful tool to understand the messages that thereby get sent among such sophisticated reasoning agents, particularly regarding deception (Kopp et al., 2018; Zaslavsky et al., 2020). This

is because it allows us to measure formally the distortion of the signals transmitted between agents, and to pin down the deception and skepticism that possibly arise. Deception thereby wears many hats. One *modus operandi* is to masquerade, for example faking confidence despite having a bad hand in Poker. Another is to completely disassociate actions from intentions, for example via a Poker face which is the same regardless of the deck.

Utility Functions in Social Interactions

Previous multiagent reinforcement learning work on shaping other agents has focused on modifying agent's utility functions to include social aspects, often based on information theory. For example, in Jaques et al. (2019) the utility function of one agent was modified to gain reward from the environment as well as from influencing other agents' behaviour. Agents endowed with this utility adopted socially aware policies, in which, for example, they learn to signal other agent on the location of an unobserved reward to lead them to otherwise unreachable locations, thus gaining much from the policy change. A similar approach was explored by Strouse et al. (2018). This work is motivated by the observation that in cooperative environments, being transparent about one's goal creates trust and coordination, while hiding one's intentions is beneficial in competitive settings. Strouse et al. (2018) handcrafted agents' utility functions to take such (in-)transparency into account: Specifically, when agents competed, their utility function was augmented with a reward for minimizing the mutual information (MI) between their actions and intentions, whereas when they were cooperated, they were rewarded for maximizing the same term.

The information-theoretic handcrafting approach is in contrast to our approach, which follows the 'reward is enough' tradition set by Silver et al. (2021). 'Reward is enough' argues that complex behaviours, particularly those observed in the social realm, do not require any handcrafting. Instead, these behaviours arise in the actions of agents that purely maximize rewards. To preview our findings, we show that this is indeed the case for deception and skepticism. We thereby crucially differ from Jaques et al. (2019) and Strouse et al. (2018) in using information theory only to measure the effects of this reward maximization but not have it be a part of the reward maximization itself. We also note that both papers, while conceptually employing ToM, do so at a shallow level, usually only endowing one agent in the environment with it, and so precluding any true recursivity.

Deception in AI

Deceptive behaviour in AI has received considerable attention, with increased interest in applications such as deceptive path planning (Masters & Sardina, 2017), military drone control (Ramirez et al., 2018) or resilience to cyber attacks (Rowe, 2003), and with a multi-agent reinforcement learning perspective (Chelarescu, 2021; Sarkadi et al., 2019). Most of this work however rests on *model-free* behaviors, where an agent learns through experience and trial-and-error to deceive. This is for example evident in ruses learnt by systems like AlphaGo (Silver et al., 2016), which uses MCTS for *Q*-value estimation but uses self-play to simulate opponent's move rather than having an explicit model of the opponent. Similarly, learning to strategically hide information, or at least not fully reveal it, can emerge despite the lack of an explicit model of the opponent, for example in the gameplay of 'Cicero' in Diplomacy (Meta Fundamental AI Research Diplomacy Team et al., 2022). However, such model-free systems are in fact sometimes easily trickable (Wang et al., 2023). Such more model-free behavior is in contrast to *model-based* (Dolan & Dayan, 2013) deception, where agents learn to target the inference processes of others, revoking the naivete of the Naive Utility Calculus.

However, existing work in this field (Aitchison et al., 2021; Liu et al., 2021) only does so at low levels of recursions, merely asking how to circumvent a naïve utility calculus but not asking how a not-so-naïve utility calculus might act, or how to act deceptively against such a savvy opponent.

METHODS

Paradigm

As an example of the emergence of disinformation, we model a buyer and a seller interacting over three periods or stages (Figure 1). Imagine a store owner offering items at stalls in various locations, and using observations of buyer behavior to set prices for subsequent sales to the buyer. We here concern ourselves with a single-shot instance of this task where the buyer and seller only interact once over these three stages, and do *not* learn about each other in a repeated fashion.

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

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

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

Approaching this task through an experimentalist's eyes, we will later vary both the distances of the two objects ($d(\text{apple})$, $d(\text{orange})$), as well as the preference the buyer has for them ($r(\text{apple})$, $r(\text{orange})$). In essence, these two, distance and preference, make this a bivariate experiment.

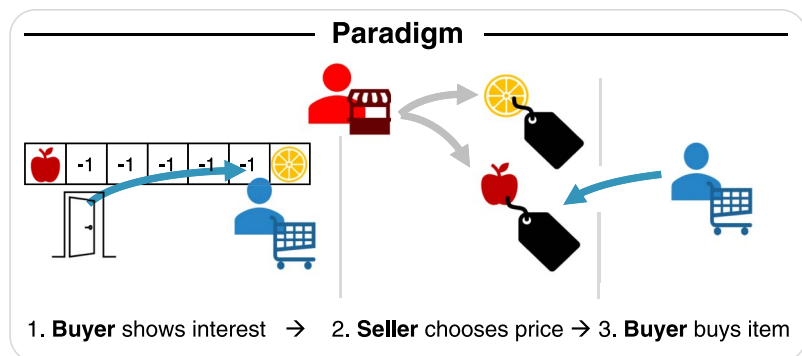


Figure 1. In our simulated paradigm, two agents interact, a buyer (blue) and a seller (red). The buyer first chooses either of two differently rewarding objects, incurring cost on the way to the chosen object via the distance travelled. The seller observes this choice and the accompanying distances, draws inferences about the buyer's prices, and then prices the items accordingly. Finally, the buyer has to buy one of the items for the set price.

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

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

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

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

$$U_B^{total}(i_1, i_3, d, m) = U_{B^1}(i_1, d(i_1)) + U_{B^3}(i_3, m(i_3)) \quad (3)$$

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

$$U_S(i_3) = m(i_3) \quad (4)$$

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

Model and Agents

We study different levels of sophistication of IPOMDP buyers and sellers arising from the recursive depth of their ToM (see Figure 2). IPOMDPs enable agents to plan through another agent's inference process via a theory of mind (Figure 1). Unlike Camerer et al. (2015), we assume a strict nesting, where each ToM-level uses only a model of the agent one step below on the ToM ladder as do Gmytrasiewicz and Doshi (2005). Recall that, at the lowest level, this can be understood as planning through another agent's inverse reinforcement learning (IRL), but can be taken further, allowing ever more sophisticated agents to model one another's inferences and planning processes recursively.

As the turns in this game alternate, the ToM levels of buyer and seller also alternate (Hula et al., 2015), starting from the simplest buyer, which we denote as ToM($k = -1$). The **ToM(-1)**

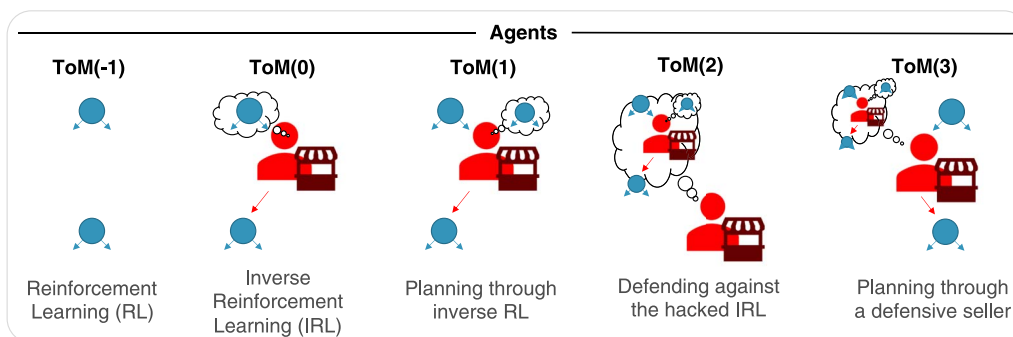


Figure 2. Model: We model agents of different levels of theory of mind (ToM). Note how all buyers and all sellers share the same respective value function and only differ in the way they use theory of mind for planning and/or inference. We begin with a simple reinforcement learning buyer in blue, ToM(-1) that makes the first and second choice independently, and a regular inverse reinforcement learning seller in red, ToM(0), who draws inferences about this buyer. The ToM(1) buyer plans through this seller's inference process to optimize its overall value. In turn, the ToM(2) seller does inverse reinforcement learning, but of a higher level, taking the ToM(1)'s planning into account. As we will see, this gives rise to a defensive inference process. The ToM(3) in turn plans just like the ToM(1) but does so with the ToM(2)'s 'defensive' inference in mind.

buyer independently decides based on distance/price and preferences at the first and third stages without taking into account the seller. That is, this naïve buyer solves what amounts to the same utility maximization problem at each step where it gets a choice. As a result, its action-values Q at a given state are simply defined as the aforementioned utilities.

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

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

After the buyer computes these Q -values, it then selects an item using a SoftMax policy (Equations 6a, 6b) with inverse temperature β . If we use j to denoting the ‘other’ item (j represents ‘apple’ if i represents ‘orange’), then the probabilities of the actions a_1 and a_3 of selecting item i_1 and i_3 at stages 1 and 3 are:

$$P_{k=-1}^{t=1}(a_1 = i_1 | \mathbf{r}, \mathbf{d}) = \sigma(\beta[Q_{k=-1}^{t=1}(i_1, d(i_1)) - Q_{k=-1}^{t=1}(j_1, d(j_1))]) \quad (6a)$$

$$P_{k=-1}^{t=3}(a_3 = i_3 | \mathbf{r}, \mathbf{m}) = \sigma(\beta[Q_{k=-1}^{t=3}(i_3, d(i_3)) - Q_{k=-1}^{t=3}(j_3, d(j_3))]) \quad (6b)$$

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

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

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

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

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

$$\begin{aligned} m_{k=0}^*(i_3, a_1) &= \arg \max_{m(i)} \{E[U_S(i) | a_1]\} \\ &= \arg \max_{m(i)} \left\{ \int m(i) \cdot \hat{P}_{k=-1}(a_3 = i | r(i), m(i)) p_{k=0}(r(i) | a_1 = i_1) dr(i) \right\} \end{aligned} \quad (8)$$

This computation states that for every possible item price $m(i)$ the seller computes the subjective probability that the ToM(−1) buyer will buy the item at that said price: $\hat{P}_{k=-1}(a_3 = i | r(i), m(i))$ using Equation 6b. This *potential* expected revenue: $\hat{U}_S(i) = m(i) \cdot \hat{P}_{k=-1}(a_3 = i | r(i), m(i))$, is weighted by the posterior probability over the buyer’s preferences $p_{k=0}(r(i) | a_1 = i_1)$, computed via the inverse reinforcement learning spelled out in Equation 7. Note that while the buyer’s decision-making is stochastic, the seller’s policy sets a deterministic price for each item.

While this has so far been a simple pairing of reinforcement learner and inverse reinforcement learner, agents with higher theory of mind will make more complex decisions.

Specifically, the **ToM(1) buyer** takes the ToM(0) price computation into account, and plans through it to optimize the sum of both first and third stage payoffs. Crucially, this model-based

planning takes into account the seller's inverse reinforcement learning, as the expectation in Equation 9 is with respect to this instance of IRL. This planning is computed through a full planning tree span. After the simulation terminates it outputs the action-values ($Q(a)$) for each item selection in the first phase. The ToM(1) selects the first action via a SoftMax policy like the ToM(-1) buyer. Since the buyer cannot affect the seller's behavior in the last step, the ToM(1) acts in an identical way to the ToM(-1), that is, it selects an item via utility maximization.

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

$$\begin{aligned} Q_{k=1}^{t=1}(a_1 = i_1, \mathbf{d}) &= U_B^{t=1}(i_1, \mathbf{d}) + E[U_B^{t=3}(i_3, m_{k=0}^*(i_3, a_1)) | a_1 = i_1] \\ &= U_B^{t=1}(i_1, \mathbf{d}) + \sum_{i_3} U_B^{t=3}(i_3, m_{k=0}^*(i_3, a_1)) \hat{P}(a_3 = i_3 | m_{k=0}^*(i_3, a_1), a_1 = i_1) \end{aligned} \quad (9)$$

Unpacking this equation, the ToM(1) agent performs a thought experiment in which it envisions itself acting as the ToM(0) seller—observing the first item pick and setting the prices accordingly. The first component of this computations is the utility from the buyer's first selection (Equation 1). Knowing that this selection (a_1) is used by the seller to determine the items' prices in the next step: $m_{k=0}^*(i_3, a_1)$, it uses its mental model of the seller to simulate the seller's optimal pricing process (Equation 8). This nested reasoning includes the seller's IRL process (Equation 7). Thus, the ToM(1) buyer can predict how its action in the first step affects the potential reward in the last phase through full mentalization of the seller's inference and learning process.

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

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

$$P_{k=1}^{t=1}(a_1 = i | \mathbf{r}, \mathbf{d}) = \sigma(\beta[Q_{k=1}^{t=1}(i_3, d(i_3)) - Q_{k=1}^{t=1}(j_3, d(j_3))]) \quad (10)$$

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

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

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

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

Information Theory, Deception and Skepticism

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

1. From the perspective of the sender of a message, we can ask *how much a message reveals* about something it wants to hide. As we will see, this will be particularly relevant when asking how much the buyer's actions reveal about its preferences. Information theory allows us to capture this using the Mutual Information between the buyer's actions and its preferences, $I(\mathbf{r}, a_1)$.
2. From the perspective of the receiver of the message, we can analyse *how much credence is lent to a signal*. We can do so by calculating how much a receiver's beliefs change in response to a signal. Information Theory lets us do this via the Kullback-Leibler (KL) Divergence between a receiver's prior and its posterior once it has seen a message. Here, we are interested in this case for the seller before and after it has seen the buyer's action: $D_{KL}(p(r|a_1)||p(r))$. Essentially, the lower this divergence, the more skeptical an agent.
3. Finally, we can take a more bird's eye view of an interaction and ask *how much a given signal sent by a sender is misinterpreted by a receiver*. Again, we can do so using KL-Divergence, for example between what the receiver assumes is a sender's policy and what the sender's actual policy is. In our case, we are for example interested in how simpler sellers might be led astray by higher level theory of mind buyers. We measure this as the KL-distance between what a ToM(k) seller *assumes* to be the buyer's policy (i.e., the ToM($k - 1$) buyer) and the ToM($k + 1$)'s actual policy, for example for the case of the ToM(0) buyer: $D_{KL}(p_{k=1}^{t=1}(a_1|r, c)||p_{k=-1}^{t=1}(a_1|r, c))$. Essentially, this quantifies how effective a sender's deception is, and is therefore different from the mutual information outlined above, which is more about the mere hiding of information (Kopp et al., 2018).

RESULTS

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

We describe the intricacies of these policies in substantial detail in order to exploit the simplicity and transparency of our setting.

We begin with the **ToM(-1) buyer**. As we outlined in the model section, this buyer acts naïvely, maximizing the utility of each stage separately. Figure 3A shows the probabilities of

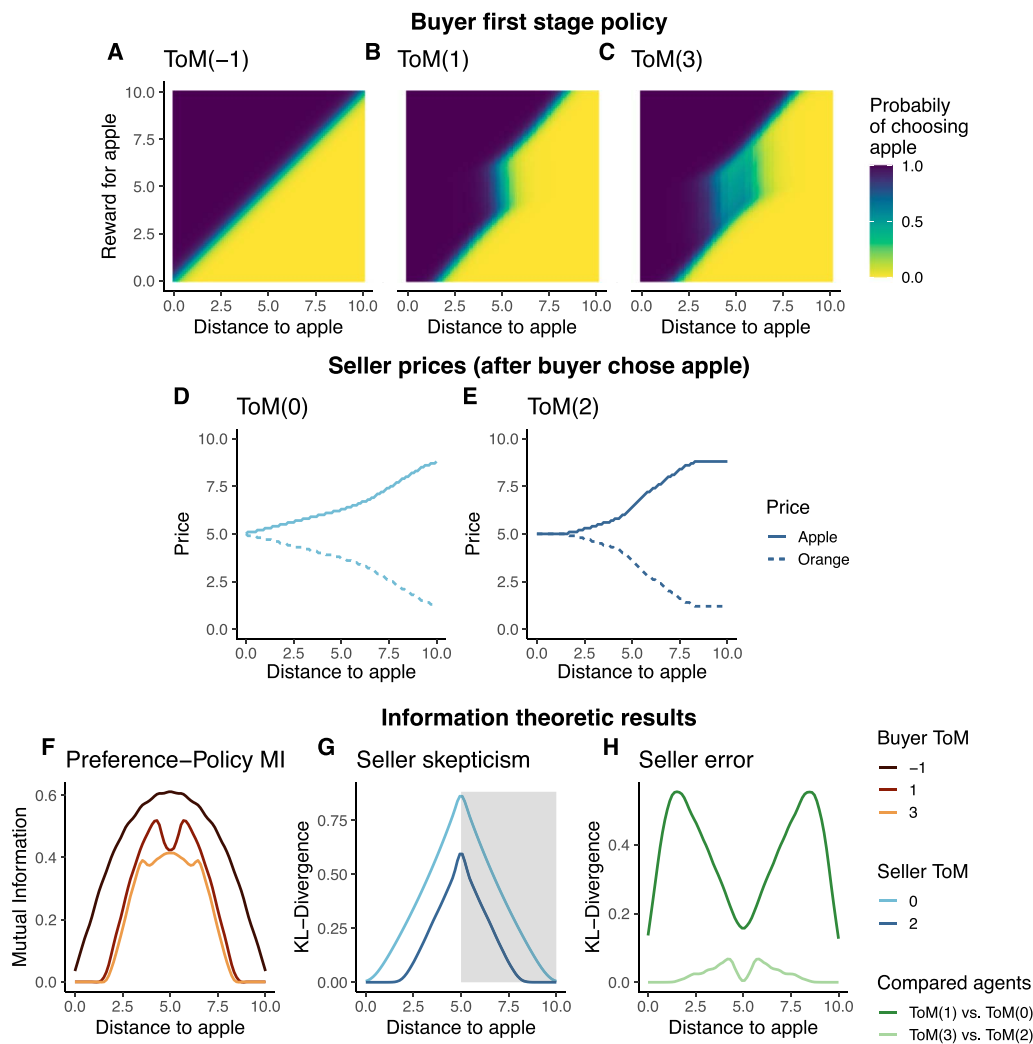


Figure 3. (A–C) Buyer policy in first stage as a function of the distance $d(\text{apple})$ and the preference towards the apple, $r(\text{apple})$, shown by different ToM-levels. Here, we are using a soft policy with a temperature of $\beta = 0.5$, as for example in equation 6a. The policies are symmetric because of the constraints. Notice the rise in ToM(1) and ToM(3), who shift their policies. (D–E) Seller prices, m , after a buyer chose the apple as a function of the distance to the apple, $d(\text{apple})$, for the two different seller ToM-levels. ToM(2) discards the evidence at the extremes. (F) Amount of deception by the different ToM buyers quantified by the mutual information between the buyer's apple preferences and the probability that they will pick an apple. ToM(1) and ToM(3) show lower values, effectively hiding their preferences. (G) Strength of the seller's belief update quantified by the KL-Divergence between their (flat) prior and posterior over the apple preferences, after observing the buyer choose the closer and thus more likely object (apple in left half, orange in shaded right half). (H) Dissimilarity between the $ToM(k)$ seller's assumed policy and the $ToM(k + 1)$ buyer's actual policy, simultaneously showing the hacking success of the buyer and error of the seller.

choosing an apple, the x-axis describing the distance from the entrance to the apple, $d(\text{apple}_1)$, and the y-axis the reward derived from consuming the apple at the end of the corridor, $r(\text{apple}_1)$. The colours represent apple selection probability. Here, due to the symmetric nature of the problem we only discuss the apple selection.

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

As we noted, can re-express such a lack of concealment in information-theoretic terms by measuring how informative the buyer's choice is about its preferences, i.e., the Mutual Information (MI) between the buyer's preferences and their initial choice, $I(\mathbf{r}, a_1)$. This MI is shown in the darkest curve in Figure 3F—the action is generally informative, particularly when the fruits are nearly equidistant from the buyer. Here, the buyer's choice reveals most about its preferences.

The ToM(−1) buyer in turn is the input to the **ToM(0) seller**. This seller can translate what it knows about the naïve ToM(−1) policy into prices via simple inverse reinforcement learning. Figure 3D shows these prices for the case that the buyer has chosen the apple in the first stage. Since the ToM(−1) signal is reliable, the distance covered by the buyer is a good proxy for its preferences. For example, if the apple is situated 8 steps from it, and the buyer picks it, then the seller can infer that the buyer's utility is at least 8 and set the price just shy of this. This all but guarantees that the apple will be selected at the last phase.

Using IT, we re-express this “trust” in the buyer's action in mathematical terms. Specifically, we measure the strength of the seller's belief update via the KL-divergence (KLD) between its (flat) prior and posterior, $D_{KL}(p(r|a_1)||p(r))$. In the lighter curve in Figure 3G, we show the ToM(0) seller's KLD associated with the choice of the more likely item, which is apple when apple is closer and orange when orange is closer (the latter shown in the shaded area). This highlights how the seller uses every decision of the buyer as a signal, irregardless of the maze set-up.

Aiming to get the best price possible, the **ToM(1) buyer** attempts to hack this pricing scheme by playing what amounts to a gambit. This manipulation is manifested in three different ways that are evident in Figure 3B. First, the ToM(1) buyer adopts a deceptive maneuver for the case when the apple is close but *undesired* (the uniformly dark-purple part between $d(\text{apple}) = 0$ and $d(\text{apple}) = 2$). In this setting, the buyer knows that the ToM(0) seller would interpret an orange selection as a signal for high orange preference. Hence it selects the apple despite its lower appeal to convince the seller that it prefers the apple to orange and gain a lower price for the orange down the road.

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

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

We can express this ruse in information-theoretic terms in two ways. Returning to the MI between (naïve) preferences and policy, we show how the ToM(1) buyer manages to significantly reduce this informativeness about its preferences, particularly when one of the items is close (the medium line in Figure 3F).

As we discussed, we can measure the success of the buyer's ruse by asking how wrong the ToM(0) seller's model is. We do so by measuring the KLD between what the ToM(0) seller *assumes* to be the buyer's policy (i.e., a ToM(−1) buyer) and the ToM(1)'s actual policy, $D_{KL}(p_{k=1}^{t=1}(a_1|r, c)||p_{k=-1}^{t=1}(a_1|r, c))$. This belief discrepancy is shown in a darker green line in Figure 3H, highlighting the large discrepancies.

Having access to this ToM(1) buyer model, the **ToM(2) seller** becomes skeptical about the buyer's actions, and adjusts its pricing appropriately. We show this pricing in Figure 3E. When the maze setting enables the ToM(1)'s bluff (for example, when the distance to the apple $d(\text{apple}_1) = 1$), observing an apple selection provides the seller with no information about the buyer's preferences (compare the previously discussed uniformly purple policy of the ToM(1) agent in Figure 3B, and notice how it always chooses the apple regardless of preference). As a result, the seller ignores the distance travelled by the buyer and keeps pricing the items equivalently at these distances.

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

Information theory again lets us formalize this skepticism via the KL-divergence as a function of the item more likely to be picked (see the dark line in Figure 3G, which shows how the ToM(2) seller's belief about the buyer is affected less, or, when the items are closer, not at all, by the buyer's actions).¹

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

This inability to outmanoeuvre the seller significantly has an information-theoretic consequence. While the Mutual Information between policy and naïve preferences of the ToM(3) buyer is, in some regions, slightly lower than the ToM(1)'s, the ToM(3) buyer cannot mischaracterize its preferences further (lightest curve in Figure 3F). Despite the limited possibilities for deception, this figure shows that through its recursive inference, the ToM(3)'s policy reduces MI, in a way that corresponds to the MI reduction of the ToM(1) agent. When the apple's distance lies in the interval $[3.0, 7.0]$, the MI between the ToM(3)'s actions and preference displays an opposite pattern relative to the ToM(1)—when the latter's MI increases, the former's MI decreases ($[3.0, 3.5]$) and when the MI of the ToM(1) decreases ($[4.0, 5.0]$), that of the ToM(3) increases. We interpret this negative correlation through the lenses of the deception of the ToM(3). When the ToM(1)'s MI increases, more information about its preferences leaks through its actions, allowing the ToM(1) seller's belief update to be more affected by the observed event, as evident in Figure 3G. Through its process of recursive inference, the

¹ We here present a case that uses flat priors throughout, particularly with regards to the initial beliefs $(p(r))$ that the seller holds about the buyer's preferences. Naturally, our paradigm would also allow for those priors to not be uniform but rather already be endowed with a first guess. Indeed, this raises the question as to whether our KL-Divergence metric is still appropriate in cases when the likelihood the seller receives aligns with its initial prior. In this case, the mean of the posterior belief of the seller would not shift. However, this reinforcement of the prior would still result in a reduction of the variance of the belief (because the seller receives an additional sample and can now be more certain about its estimate). In this case, the KL-Distance $D_{KL}(p(r|a_1)||p(r))$ would still be above 0, indicating that the seller takes the buyers actions into consideration. In contrast, when the seller knows (through theory of mind), that the buyer is sending an uninformative signal, then it would simply keep its prior mean and variance, resulting in $D_{KL}(p(r|a_1)||p(r)) = 0$, just as we showed before—even when the likelihood might superficially 'agree' with the prior to a naïve receiver.

ToM(3) buyer determines that it has to reduce the MI to avoid preference detection by the ToM(2) seller. In a less constrained environment, this antithetical pattern may be stronger. Equally, the discrepancy between the ToM(2) seller's assumptions about the ToM(3) buyer and the truth is much less than that for the ToM(0) seller and ToM(1) buyer pair (light curve in Figure 3H). Note that the ToM(2) dissimilarity increases in regions where the ToM(0) dissimilarity decreases, showing the ToM(3)'s attempts at deception.

DISCUSSION AND FUTURE WORK

Our work shows how purely reward-maximizing agents can appear to engage in complex signalling behavior, as captured by information theory. Unlike Strouse et al. (2018), where the mutual information manipulation is pre-programmed into the utility function of the agents, in our work we only rely on theory of mind and planning. We also extend multi-agent reinforcement learning work on deception (Aitchison et al., 2021; Liu et al., 2021) deeper into the cognitive hierarchy. We show how, through a form of planning that is opponent-aware, agents can exploit other agents' inference processes. More specifically, we show how a sender can purposefully reduce the informativeness of their actions and target the inferences they expect a receiver to perform. On the other hand, we show how agents can defend themselves from this manipulation by (partially) ignoring the behaviour they observe. This counter-deception can be interpreted as skepticism and we illustrate it both in policies and in inference.

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

This work is relevant for the study of social cognition in artificial (Jaques et al., 2019; Rabinowitz et al., 2018) and biological systems. For example, it adds a model-based reinforcement learning and information theoretic perspective to Goodhart's law (Goodhart, 1984), which states that people tend to try to game statistical regularities used by authorities (e.g., the government or a retailer) for control purposes (e.g., taxation or dynamic pricing). Our agents also engage in rational 'deliberate ignorance', or the practice of purposefully discarding signals that they might receive. Deliberate ignorance has received interest as both a description of empirical human behavior and as a potentially useful strategy in game-theory and law (Hertwig & Engel, 2016, 2021). Future research will have to investigate the crucial question of how closely humans (Barnett et al., 2021; Ransom et al., 2017), or other animals (Premack & Woodruff, 1978), actually follow our theoretic analyses in this task.

In general, animals and humans can engage in what appears to be behavior that is similar to what we observe theoretically. For example, Clayton et al. (2007) and Emery and Clayton (2001) observed that corvid engage in sophisticated ruses to hide their food-cache when they are being watched. Of particular interest here is that only those corvids that had previously stolen do so, raising interesting questions about how biological agents learn to rise in the ToM hierarchy.

In humans, ToM-driven deception has also been observed. For example, Crawford (2003) analyzed the Allies' decision to land in Normandy instead of Calais through the lens of theory of mind. Relevant to our work, Oey et al. (2023) use a recursive modelling framework similar to ours. They show that only a recursive theory of mind style model could explain several empirical facets of how senders design lies. Using less formal models, the developmental

psychology literature has investigated how theory of mind may be a precursor for children's ability to lie: For example Ding et al. (2015) show how teaching kids theory of mind skills led them to 'deceive constantly', and in two larger scale meta-analyses Sai et al. (2021) and Lee and Imuta (2021) show a correlation between children's theory of mind abilities and lying. Seeing that inverse reinforcement learning has been taken as a basis for benevolence about other people in the form of the Naïve Utility Calculus, we believe that IPOMDP, framed as an extension of IRL, is a strong candidate for extending this reasoning systematically to more competitive domains. Thereby, IPOMDP allow for a joint Bayesian reinforcement learning account of social interaction (FeldmanHall & Nassar, 2021; FeldmanHall & Shenhav, 2019). We also suggest that information theory will be a useful tool to understand the messages sent between deceptive and skeptical children, and adults.

Theory of Mind: Advantages and Limitations

Our work focuses on deep recursion in theory of mind. The benefits of such recursion have been studied in multiple environments, ranging from cooperative (Devaine et al., 2014) to competitive (Hula et al., 2015). While it is well accepted that higher ToM level can improve the outcome of the individual, it is not clear whether this improvement is linear or asymptotic. Our work speaks to this by showing how a first step of deception (ToM(1)) is countered with relatively sweeping skepticism (ToM(2)). This in turn locks down the ability for further deception in higher ToM(3) evident in Figure 3H), hinting that, at least in settings similar to ours, higher ToM benefits are asymptotic. We note that these dynamics are highly affected by the conservative pricing policy adopted by the ToM(2) seller as a precautionary mechanism against the ToM(1) deception. In turn, the ToM(3) is cornered, hence its policy resembles the ToM(1)'s policy.

One crucial limitation for the biological plausibility of our simulations is the high computational cost of deep recursive reasoning. This recursion forces a ToM(k) agent to solve IM^k POMDP problems where M is the number of possible agent models. In Gmytrasiewicz and Doshi (2005) this complexity is evaluated to be PSPACE-hard for finite time horizon. From both computational complexity as well as from a bounded rationality perspective, this implies that the optimal policy of higher ToM levels is practically infeasible to compute in an accurate and timely manner. Indeed, previous theoretical work limited the highest ToM level to, in our nomenclature ToM(2) (for example, Adhikari & Gmytrasiewicz, 2021; Devaine et al., 2014; Goodie et al., 2012). Future work on efficient recursive beliefs representation is needed to make this problem computationally feasible.

While we highlighted the benefits of higher ToM levels, high ToM can also be detrimental. Of particular interest are misplaced high levels of recursion that may lead agents to be unnecessarily skeptical, when they are in fact faced with naïve agents. This would arise from our k -level cognitive hierarchy model where ToM(k) always strictly interpret actions as coming from the less sophisticated agent that is specifically ToM($k - 1$). Alon et al. (2023) investigated equivalent IPOMDP agents in a dynamic iterated game, showing deception and skepticism similar to the present work, but showing how misplaced high theory of mind gives rise to agents that stop trusting each other and lose reward in the process.

Further Work

Our work leaves much opportunity for further investigation: First, a key factor of these deceptive dynamics is full observability of the actions of the players. Hence, future work might explore the information-seeking perspective of this model (Schulz et al., 2023), asking how

much the buyer is willing to pay, or be paid, to disclose its actions and how much the seller is willing to pay to uncover the buyer's action.

Secondly, certain settings of our task—like the extreme corners of the maze—encourage more or less deceptive behaviours. Future work might thus treat the settings of the task as an endogenous variable that the seller can control as part of its utility maximization planning. Naturally, in such a setting, buyers should have the option to avoid the shopping task altogether. We note that the control of the decision environment has links to the game theoretic work on preference elicitation and mechanism design (Becker et al., 1964; Roth, 2008).

Finally, potential future work could adopt a macroeconomic perspective and explore the multi-seller, multi-buyer case. In this setting, sellers need to make inferences about the actions of their competitors as well as about the preferences of their clients. In turn, the buyers' planning might include a more sophisticated price query policy.

Ethical and AI Alignment Concerns

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

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

A crucial aspect of the safety and ethics perspective on our results is that it is not theory of mind alone that gives rise to deception and skepticism. Rather, what produces this behavior is the coupling of theory of mind with a value function that mis-aligns the utilities of two agents. Our work therefore speaks to the alignment problem in AI safety. For example, the deceitful hiding of intentions may be crucial for the off-switch game (Hadfield-Menell et al., 2017).

ACKNOWLEDGMENTS

We would like to thank Rahul Bhui and Stefan Bucher for helpful comments. This research was been partly funded by Israel Science Foundation grant #1340/18 (NA; JSR), 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.

REFERENCES

- Adhikari, S., & Gmytrasiewicz, P. J. (2021). Telling friend from foe—Towards a Bayesian approach to sincerity and deception. In D. Wang, R. Falcone & J. Zhang (Eds.), *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. CEUR-WS.org
- Aitchison, M., Benke, L., & Sweetser, P. (2021). Learning to deceive in multi-agent hidden role games. In S. Sarkadi, B. Wright, P. Masters, & P. McBurney (Eds.), *Deceptive AI, communications in computer and information science* (pp. 55–75). Springer International Publishing. https://doi.org/10.1007/978-3-030-91779-1_5
- Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with large language models. *arXiv:2305.16867*. <https://doi.org/10.48550/arXiv.2305.16867>
- Alon, N., Schulz, L., Dayan, P., & Barnby, J. M. (2023). Between prudence and paranoia: Theory of mind gone right, and wrong. In *First workshop on theory of mind in communicating agents*. Retrieved from <https://openreview.net/forum?id=gB9zrEjhZD>.
- Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 33, pp. 2469–2474). Cognitive Science Society.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 1–10. <https://doi.org/10.1038/s41562-017-0064>
- Barnett, S. A., Hawkins, R. D., & Griffiths, T. L. (2021). A pragmatic account of the weak evidence effect. *arXiv: 2112.03799*. <https://doi.org/10.48550/arXiv.2112.03799>
- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3), 226–232. <https://doi.org/10.1002/bs.3830090304>, PubMed: 5888778
- Berke, M., & Jara-Ettinger, J. (2021). Thinking about thinking through inverse reasoning. *PsyArXiv*. <https://doi.org/10.31234/osf.io/r25qn>
- Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3), 861–898. <https://doi.org/10.1162/0033553041502225>
- Camerer, C. F., Ho, T.-H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3, 157–162. <https://doi.org/10.1016/j.cobeha.2015.04.005>
- Chelarescu, P. (2021). Deception in social learning: A multi-agent reinforcement learning perspective. *arXiv:2106.05402*. <https://doi.org/10.48550/arXiv.2106.05402>
- Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences*, 362(1480), 507–522. <https://doi.org/10.1098/rstb.2006.1992>, PubMed: 17309867
- Crawford, V. P. (2003). Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 93(1), 133–149. <https://doi.org/10.1257/00028280321455197>
- De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6), 1222–1231. <https://doi.org/10.1016/j.neuron.2013.07.003>, PubMed: 24050407
- Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, 10(12), Article e1003992. <https://doi.org/10.1371/journal.pcbi.1003992>, PubMed: 25474637
- Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science*, 26(11), 1812–1821. <https://doi.org/10.1177/0956797615604628>, PubMed: 26431737
- Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2), 312–325. <https://doi.org/10.1016/j.neuron.2013.09.007>, PubMed: 24139036
- Doshi, P., Gmytrasiewicz, P., & Durfee, E. (2020). Recursively modeling other agents for decision making: A research perspective. *Artificial Intelligence*, 279, Article 103202. <https://doi.org/10.1016/j.artint.2019.103202>
- Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, 414(6862), 443–446. <https://doi.org/10.1038/35106560>, PubMed: 11719804
- FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, 25(12), 1045–1057. <https://doi.org/10.1016/j.tics.2021.09.002>, PubMed: 34583876
- FeldmanHall, O., & Shenav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, 3(5), 426–435. <https://doi.org/10.1038/s41562-019-0590-x>, PubMed: 31011164
- Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24, 49–79. <https://doi.org/10.1613/jair.1579>
- Goodhart, C. A. (1984). Problems of monetary management: The UK experience. In *Monetary theory and practice* (pp. 91–121). Springer. https://doi.org/10.1007/978-1-349-17295-5_4
- Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1), 95–108. <https://doi.org/10.1002/bdm.717>
- Guo, F. (2023). GPT agents in game theory experiments. *arXiv:2305.05516*. <https://doi.org/10.48550/arXiv.2305.05516>
- Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. *arXiv:1611.08219*. <https://doi.org/10.48550/arXiv.1611.08219>
- Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, 11(3), 359–372. <https://doi.org/10.1177/1745691616635594>, PubMed: 27217249
- Hertwig, R., & Engel, C. (2021). *Deliberate ignorance: Choosing not to know*. MIT Press. <https://doi.org/10.7551/mitpress/13757.001.0001>
- Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11), 959–971. <https://doi.org/10.1016/j.tics.2022.08.003>, PubMed: 36089494
- Hula, A., Montague, P. R., & Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, 11(6), Article e1004254. <https://doi.org/10.1371/journal.pcbi.1004254>, PubMed: 26053429
- Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., & De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning* (pp. 3040–3049). PMLR.

- Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29, 105–110. <https://doi.org/10.1016/j.cobeha.2019.04.010>
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8), 589–604. <https://doi.org/10.1016/j.tics.2016.05.011>, PubMed: 27388875
- Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, 140, 14–23. <https://doi.org/10.1016/j.cognition.2015.03.006>, PubMed: 25867996
- Kopp, C., Korb, K. B., & Mills, B. I. (2018). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to “fake news”. *PLoS One*, 13(11), Article e0207383. <https://doi.org/10.1371/journal.pone.0207383>, PubMed: 30485356
- Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv:2302.02083*. <https://doi.org/10.48550/arXiv.2302.02083>
- Lee, J. Y. S., & Imuta, K. (2021). Lying and theory of mind: A meta-analysis. *Child Development*, 92(2), 536–553. <https://doi.org/10.1111/cdev.13535>, PubMed: 33462865
- Liu, Z., Yang, Y., Miller, T., & Masters, P. (2021). Deceptive reinforcement learning for privacy-preserving planning. *arXiv:2102.03022*. <https://doi.org/10.48550/arXiv.2102.03022>
- Masters, P., & Sardina, S. (2017). Deceptive path-planning. In *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 4368–4375). <https://doi.org/10.24963/ijcai.2017/610>
- Meta Fundamental AI Research Diplomacy Team, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., ... Zijlstra M. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067–1074. <https://doi.org/10.1126/science.ade9097>, PubMed: 36413172
- Miura, S., & Zilberstein, S. (2021). A unifying framework for observer-aware planning and its complexity. In C. de Campos & M. H. Maathuis (Eds.), *Proceedings of the 37th conference on uncertainty in artificial intelligence* (Vol. 161, pp. 610–620). PMLR.
- Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of 17th International conference on machine learning* (pp. 663–670). Morgan Kaufmann.
- Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 152(2), 346–362. <https://doi.org/10.1037/xge0001277>, PubMed: 35980710
- Oguntola, I., Hughes, D., & Sycara, K. (2021). Deep interpretable models of theory of mind. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)* (pp. 657–664). IEEE. <https://doi.org/10.1109/RO-MAN50785.2021.9515505>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227). PMLR.
- Ramachandran, D., & Amir, E. (2007). Bayesian inverse reinforcement learning. In *Proceedings of the 20th International joint conference on artificial intelligence, IJCAI'07* (pp. 2586–2591). Morgan Kaufmann Publishers Inc.
- Ramirez, M., Papasimeon, M., Lipovetzky, N., Benke, L., Miller, T., Pearce, A. R., Scala, E., & Zamani, M. (2018). Integrated hybrid planning and programmed control for real time UAV maneuvering. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 1318–1326). International Foundation for Autonomous Agents and Multiagent Systems.
- Ransom, K., Voorspoels, W., Perfors, A., & Navarro, D. (2017). A cognitive analysis of deception without lying. In *Proceedings of the 39th annual meeting of the cognitive science society*. (pp. 992–997). Cognitive Science Society.
- Ray, D., King-Casas, B., Montague, P., & Dayan, P. (2008). Bayesian model of behaviour in economic games. In *Proceedings of the 21st international conference on neural information processing systems* (pp. 1345–1352). Curran Associates Inc.
- Roth, A. E. (2008). Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory*, 36(3), 537–569. <https://doi.org/10.1007/s00182-008-0117-6>
- Rowe, N. C. (2003). Counterplanning deceptions to foil cyber-attack plans. In *IEEE systems, man and cybernetics society information assurance workshop* (pp. 203–210). IEEE. <https://doi.org/10.1109/SMCSIA.2003.1232422>
- Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., & Gläscher, J. (2020). Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146, Article 107488. <https://doi.org/10.1016/j.neuropsychologia.2020.107488>, PubMed: 32407906
- Sai, L., Shang, S., Tay, C., Liu, X., Sheng, T., Fu, G., Ding, X. P., & Lee, K. (2021). Theory of mind, executive function, and lying in children: A meta-analysis. *Developmental Science*, 24(5), Article e13096. <https://doi.org/10.1111/desc.13096>, PubMed: 33544950
- Sap, M., LeBras, R., Fried, D., & Choi, Y. (2023). Neural theory-of-mind? On the limits of social intelligence in large LMs. *arXiv:2210.13312*. <https://doi.org/10.48550/arXiv.2210.13312>
- Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., & Chapman, M. (2019). Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4), 287–302. <https://doi.org/10.3233/AIC-190615>
- Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3), 604–639. <https://doi.org/10.1037/rev0000401>, PubMed: 36757948
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>, PubMed: 26819042
- Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, 299, Article 103535. <https://doi.org/10.1016/j.artint.2021.103535>
- Steixner-Kumar, S., Rusch, T., Doshi, P., Spezio, M., & Gläscher, J. (2022). Humans depart from optimal computational models of interactive decision-making during competition under partial information. *Scientific Reports*, 12(1), Article 289. <https://doi.org/10.1038/s41598-021-04272-x>, PubMed: 34997138
- Strouse, D., Kleiman-Weiner, M., Tenenbaum, J., Botvinick, M., & Schwab, D. J. (2018). Learning to share and hide intentions using information regularization. In *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.

- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv:2302.08399*. <https://doi.org/10.48550/arXiv.2302.08399>
- Wang, T. T., Gleave, A., Tseng, T., Belrose, N., Pelrine, K., Miller, J., Dennis, M. D., Duan, Y., Pogrebniak, V., Levine, S., & Russell, S. (2023). Adversarial policies beat superhuman go Als. *arXiv:2211.00241*. <https://doi.org/10.48550/arXiv.2211.00241>
- Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, 8(12), Article e1002841. <https://doi.org/10.1371/journal.pcbi.1002841>, PubMed: 23300423
- Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, 152(6), 1565–1579. <https://doi.org/10.1037/xge0001336>, PubMed: 36877460
- Zaslavsky, N., Hu, J., & Levy, R. P. (2020). A rate-distortion view of human pragmatic reasoning. *arXiv:2005.06641*. <https://doi.org/10.48550/arXiv.2005.06641>

Chapter 3

Paper II: (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia

Published as: Alon, N., Schulz, L., Bell, V., Moutoussis, M., Dayan, P., & Barnby, J.M. (2024). (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1), 159–177. <https://doi.org/10.5334/cpsy.117>

Author contributions: N.A. led data curation, formal analysis, investigation, methodology, software, and original draft preparation. L.S. contributed formal analysis and visualisation. V.B. and M.M. contributed to writing (review & editing). P.D. and J.M.B. contributed to conceptualisation, supervision, and writing.



(Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia

RESEARCH ARTICLE

NITAY ALON 

LION SCHULZ 

VAUGHAN BELL 

MICHAEL MOUTOUSSIS 

PETER DAYAN 

JOSEPH M. BARNBY 

*Author affiliations can be found in the back matter of this article

 ubiquity press

ABSTRACT

Humans need to be on their toes when interacting with competitive others to avoid being taken advantage of. 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 mentalizing 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 mentalizing too deeply – hyper-mentalizing – false beliefs arise that a partner is trying to trick them maliciously, resulting in a material loss to the self. Importantly, we show that this is only true when hypermentalizing agents believe observed actions are generated intentionally. 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.

AUTHOR SUMMARY

Interacting competitively requires vigilance to avoid deception. However, excessive caution can have adverse effects, stemming from false beliefs of intentional harm. So far there is no formal cognitive account of what may cause this suspiciousness. Here we present an examination of this phenomenon through the lens of Theory of Mind – the cognitive ability to consider the beliefs, intentions, and desires of others. By simulating interacting computer agents we illustrate how well-aligned agents can give rise to successful deception and justified skepticism. Crucially, we also reveal that overly cautious agents develop false beliefs that an ingenuous partner is attempting malicious trickery, leading to tangible losses. As well as formally defining a plausible mechanism for suspiciousness, paranoia, and conspiratorial thinking, our theory indicates that rather than a deficit in Theory of Mind, paranoia may involve an over-application of strategy to genuine behaviour.

CORRESPONDING AUTHOR:

Nitay Alon

Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel; Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
nitay.alon@mail.huji.ac.il

KEYWORDS:

Theory of mind; paranoid delusions; paranoia; emergence of paranoid behaviour; computational psychiatry; computational model; Bayesian brain

TO CITE THIS ARTICLE:

Alon, N., Schulz, L., Bell, V., Moutoussis, M., Dayan, P., & Barnby, J. M. (2024). (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1), pp. 159–177. DOI: <https://doi.org/10.5334/cpsy.117>

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 (Dennett, 1989; Premack & Woodruff, 1978).

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) (Barnby, Bellucci, et al., 2023; Ray, King-Casas, Montague, & Dayan, 2008; C. Baker, Saxe, & Tenenbaum, 2011; C. Baker & Tenenbaum, 2014; C. L. Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Goodman et al., 2006). 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 (Ng & Russell, 2000; Jara-Ettinger, 2019). 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 (Camerer, Ho, & Chong, 2004; O’Grady, Kliesch, Smith, & Scott-Phillips, 2015) – has been suggested as supporting the way that humans choose what to say or teach to maximise interpretability (Goodman & Frank, 2016; Barnett, Griffiths, & Hawkins, 2022), and as underlying cognition in social, competitive settings (Devaine, Hollard, & Daunizeau, 2014a). 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 (Alon, Schulz, Rosenschein, & Dayan, 2023; Doshi, Qu, & Goodie, 2014).

With ToM’s outsized role in human interaction (Devaine, Hollard, & Daunizeau, 2014b), it is unsurprising that failures of ToM have been suggested as being at least part of the basis of several psychiatric disorders (McLaren, Gallagher, Hopwood, & Sharp, 2022), such as autism (Frith & Happé, 1994; Yoshida et al., 2010; Chiu et al., 2008), psychosis (Bentall & Kinderman, 1998; Randall, Corcoran, Day, & Bentall, 2003; Penn, Sanna, & Roberts, 2008), and personality disorders (Sharp et al., 2011; Hula, Montague, & Dayan, 2015a; Galvez-Merlin et al., 2023; Rifkin-Zybutz et al., 2021; Euler et al., 2021; King-Casas et al., 2008).

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 (Buck, Browne, Gagen, & Penn, 2023). In borderline personality disorder (BPD), individuals are theorised as attributing an excessively high level of intentionality to sparse social data (Sharp et al., 2011). Here, over-mentalizing or hyper-mentalizing is defined as “making excessively convoluted inferences based on others’ social cues” and (Fonagy, Luyten, & Bateman, 2015) has been suggested as giving rise to paranoia in BPD (McLaren, Gallagher, Hopwood, & Sharp, 2022), and was shown empirically to be related to early stages of disorder (Galvez-Merlin et al., 2023). In both psychosis and BPD, and even more commonly in conspiratorial ideation (Bowes, Costello, & Tasimi, 2023), 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 (Penn et al., 2008; Bell, Mills, Modinos, & Wilkinson, 2017). Traditionally ToM has been measured using vignette tasks. While these are typically predictive of psychosis (Fett et al., 2011), and to some extent persecutory delusions (Corcoran et al., 2007), a major limitation is that they do not take into consideration the self-representation of the participants, nor the representation of others about their sense of self (Chan & Chen, 2011). Some computational and experimental work using game theory paradigms suggests that high paranoia and psychosis are affiliated with rigid, slower-to-update priors about the self-relevant nature of a partner’s actions (Fett et al., 2012; Barnby, Mehta, & Moutoussis, 2022; Barnby, Bell, Deeley, Mehta, & Moutoussis, 2023). This begs the question as to whether ToM changes in paranoia and paranoia-affiliated diagnoses may be caused by changes in the maladaptive application of recursive cognition in social settings. There has been little work examining the role of cognitive recursion applied to BPD, paranoia, and persecutory delusions,

aside from some notable exceptions (Hula, Vilares, Lohrenz, Dayan, & Montague, 2018; Hula et al., 2015a), 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 (IPOMDP) (Gmytrasiewicz & Doshi, 2005) 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 (Ho, Saxe, & Cushman, 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: 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 mentalizing in mixed-motive games. These sequential social dilemmas (SSD) serve as a tractable testbed to observe the emergence of complex behaviour (Alon, Schulz, Rosenschein, & Dayan, 2023), 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 Mentalization (DoM) can successfully manipulate the beliefs of those one-step lower in the hierarchy (Alon, Schulz, Rosenschein, & Dayan, 2023; Alon, Schulz, Dayan, & Barnby, 2023).

Next, we present the potential downside associated with maladaptively high DoM, i.e., hyper- or over-mentalizing. 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.

DEFINITION: OVERMENTALIZING

When properly calibrated, Theory of Mind (ToM) is used to reason about the mental state of others, inferring their intentions from actions, and thus being able to respond or anticipate appropriately in the future. In formal terms has been framed as a hierarchical system: a $\text{DoM}(k)$ agent properly models a $\text{DoM}(k - 1)$ agent and utilizes this ability to predict and affect the behaviour of that $\text{DoM}(k - 1)$ agent.

This fixed hierarchy assumes that a self can model others one step below. What has not been traditionally considered is what happens when we infer sophisticated or complex intentions to otherwise simple or ingenuous behaviour. To put this into context, imagine walking in the street and seeing a friend. You wave, but your friend does not wave back. The reality is that your friend has not seen you, but this is unknown to you. A simple interpretation is to assume they did not see you and therefore did not wave. A more recursive hypothesis is that your friend did see you, but chose not to wave. This interpretation assumes your friend has a more complex model of the situation, one that involves you and a conscious choice to ignore you. Lastly, one can imagine an even deeper recursion, one in which you assume that not only your (so called) friend saw you and decided to ignore you, but that wasn't by accident that you two met on the street – rather they followed you to that street, planning to cause you to be upset by pretending to not see you. Formally, we can think of our friend as a $\text{DoM}(k = -1)$ agent in this context – they are making (or omitting actions) without any regard to you. The first interpretation is an example of a $\text{DoM}(k = 0)$ belief – one that is focused on the other with no regard to the self. The second interpretation uses a $\text{DoM}(k = 1)$ model, which means that your friend is thinking about you at the time when they omitted to wave. Thus, as a general rule, when the observed agent's DoM is lower than $(k - 1)$, the $\text{DoM}(k)$ agent overestimates the complexity of the observer's world model, which leads to an overly sophisticated internal simulation of potential intentions that accounted for their actions. This, in essence, is overmentalizing.

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 veridicality and reward between agents. As a result, our work has key implications for AI safety and human-computer interaction.

MATERIALS AND METHODS

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 are competing or mixed 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 (Xiang, Lohrenz, & Montague, 2013; Alon, Schulz, Dayan, & Barnby, 2023) IUG (Figure 1). This game is comprised of $T > 0$ repetitions of the following game: a sender, S , is endowed with 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.

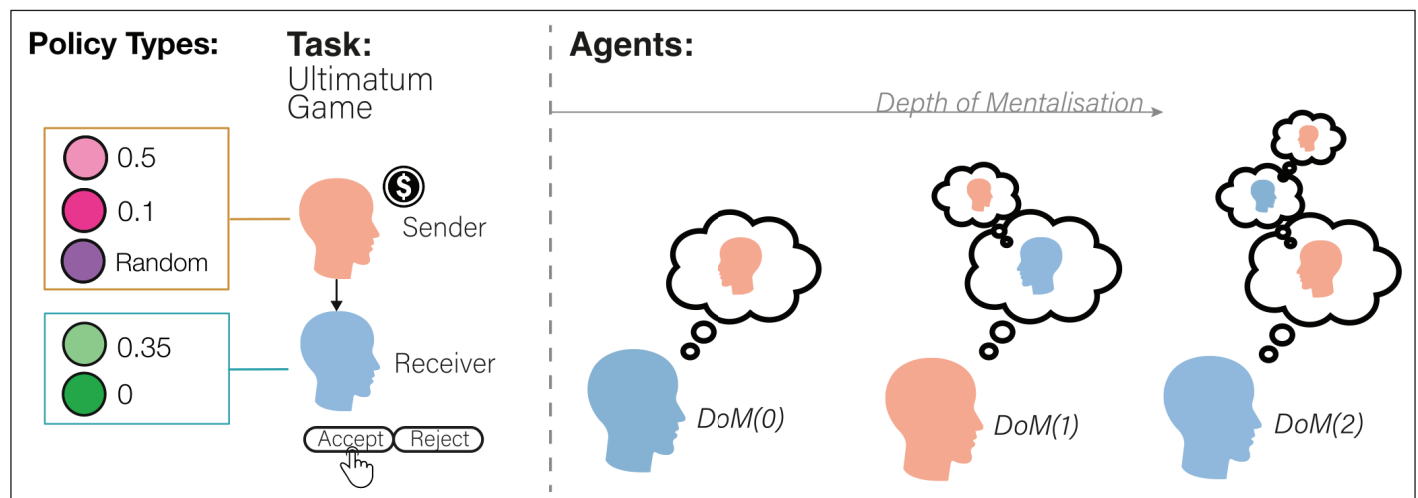


Figure 1 Task and Agent Summary: In the Ultimatum Game, a sender (orange) chooses how much of an endowment to send to a receiver (blue). The receiver then has a chance to either accept or reject this offer. If the receiver accepts, they both get to keep their portion of the endowment. If the receiver 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 Mentalization 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. Agents at $DoM(>0)$ can conceptualise the world model of how others below them in the hierarchy perceive the self. Given the requirement of the IUG to balance short and long-term rewards, and given the types of agents available, this leads to strategic play by more sophisticated agents to get a greater frequency of more favourable outcomes. For example, a $DoM(1)$ agent knows that the $DoM(0)$ is able to conceptualise whether they are playing with a random or intentional sender, and therefore may behave in a way that causes the $DoM(0)$ to mischaracterise the sender's identity as random. This means the $DoM(1)$ can then send very unfavourable offers knowing that the $DoM(0)$ does not believe they can influence the outcome.

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: $h^t = \langle a_S^1, a_R^1, \dots, a_S^t, a_R^t \rangle$.

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 $\eta_S \geq 0$ and $\eta_R \geq 0$, 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, represented 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 (Fehr & Schmidt, 1999) adds to the agent's utility gain (loss) from inequality aversion. We keep this option for future research.

Formally, the utilities of agents with thresholds η_S, η_R are:

$$u_S^t(\eta_S, a_S^t, a_R^t) = (1 - a_S^t - \eta_S) * a_R^t \quad (1)$$

$$u_R^t(\eta_R, a_S^t, a_R^t) = (a_S^t - \eta_R) * a_R^t \quad (2)$$

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 T :

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

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

We model the agents using the IPOMDP framework (Gmytrasiewicz & Doshi, 2005). This framework augments the POMDP model to account for modelling 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 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 (Hula, Montague, & Dayan, 2015b) – senders and receivers have odd- and even-numbered DoM respectively.

At the bottom of the hierarchy are DoM(-1) (sub-intentional) agents. DoM(-1) agents are characterized by lacking an opponent model (belief about the other) 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 and test strategies used to exploit the possible existence of a random other. This is both useful to examine when an agent may use the presence of a random other to their advantage (acting as a random agent to instil a sense of powerlessness in their opponent) and thus when an agent's applied strategy may mistake randomness for intentional policy.

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:

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

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

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

$$Q_{S=-1}^t(a_s^t; \eta_s) = u_s^t(a_s^t, \eta_s) \quad (6)$$

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 (Ng & Russell, 2000). Upon observing an offer a_s^t , the DoM(0) receiver computes the likelihood of the offer for each possible sender type and re-weights them with current beliefs:

$$b_{R=0}^t(\hat{\theta}_{S=-1}) = p_{R=0}^t(\hat{\theta}_{S=-1} | h^{t-1}, a_s^t) \propto p_{S=-1}^t(a_s^t | h^{t-1}, \hat{\theta}_{S=-1}) b_{R=0}^{t-1}(\hat{\theta}_{S=-1}) \quad (7)$$

where, $p_{S=-1}^t(a_s^t | h^{t-1}, \hat{\theta})$ 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 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 :

$$Q_{R=0}^t(a_R^t; \eta_R, b_{R=0}^t(\theta_{S=-1})) = E_{a_s^{t+1} \sim \pi_{S=-1}^*} \left[u_R^t(a_s^t, \eta_R) \cdot a_R^t + \gamma \max_{a_R^{t+1}} \{ Q_{R=0}^{t+1}(a_R^{t+1}; \eta_R, b_{R=0}^{t+1}(\theta_{S=-1})) \} \right] \quad (8)$$

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 (Hutter, 2004). 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 includes beliefs about the receiver's threshold and recursively 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:

$$b_{S=1}^t(\hat{\theta}_{R=0}, b_{R=0}^{t-1}(\hat{\theta}_{S=-1})) = p_{S=1}^t(\hat{\theta}_{R=0}, b_{R=0}^{t-1} | h^{t-1}) \propto p_{R=0}^t(a_R^{t-1} | h^{t-2}, a_s^{t-1}, \hat{\theta}_{R=0}, b_{R=0}^{t-1}(\hat{\theta}_{S=-1})) b_{S=1}^{t-1}(\hat{\theta}_{R=0}, b_{R=0}^{t-1}(\hat{\theta}_{S=-1})) \quad (9)$$

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:

$$Q_{S=1}^t(a_s^t; \eta_s, b_{S=1}^t(\hat{\theta}_{R=0}, b_{R=0}^{t-1})) = E_{a_R^t \sim \pi_{R=0}^*} \left[u_s^t(a_s^t, \eta_s) \cdot a_s^t + \gamma \max_{a_s^{t+1}} \{ Q_{S=1}^{t+1}(a_s^{t+1}; \eta_s, b_{S=1}^{t+1}(\hat{\theta}_{R=0}, b_{R=0}^t)) \} \right] \quad (10)$$

Much like the DoM(0) receiver, the DoM(1) sender also uses internal simulation to infer how its actions will affect the receiver. However, while the DoM(0) receiver can manipulate the bounds of the DoM(-1) sender, the DoM(1) sender has a representation of this, and thus can anticipate and predict the beliefs of the DoM(0) receiver to their favour.

We also consider a DoM(2) receiver. This agent models the sender as DoM(1), including all DoM(1) nested beliefs. These nested beliefs are the beliefs the DoM(1) sender ascribes to the presumed DoM(0) receiver as mentioned above. Notably, these receivers also consider the random sender in their model, given that the DoM(1) agent uses the existence of a random agent as a mechanism to deceive the DoM(0). 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 (Hula et al., 2015b), 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. A summary of the agents properties is presented in Table 1.

	DoM(-1)	DoM(0)	DoM(1)	DoM(2)
DoM characterisation	None	Infer direct action about the other irrespective of the self	Consider others' beliefs about the self	Consider other' beliefs about the self thinking about the other
Formalism	None	$b_{k=0}(\theta) = P(\eta_{k=-1})$	$b_{k=1}(\theta) = P(\eta_{k=0}) \times b_{k=0}(\theta)$	$b_{k=2}(\theta) = P(\eta_{k=1}) \times b_{k=1}(\theta)$
Strategic behaviour	Reactive	Identification of the other based on their history of actions	Belief manipulation based on an other's model (k-1) of the self	Skepticism, Counter-deceptive knowledge that a DoM(1) agent will try and deceive
Task behaviour	Responds to action based on a fixed policy	Adapted behaviour to observations given the identified opponent type	Selects actions to manipulate beliefs given the known world model of a DoM(0) agent	Identify deceptive behaviour and counter-deceive DoM(1)
Vulnerability	DoM(0) manipulation	DoM(1) manipulation	DoM(2) counter-deceptive manipulation	Overmentalizing about DoM(0) and DoM(-1)

Table 1 Summary of DoM levels and behaviours. Here we refer to 'self' to mean the subject of the belief, and 'other' to mean the partner.

In this work we set the experiment parameters to be $T = 12$ and two sender thresholds: $\eta_s \in \{0.1, 0.5\}$ and two receiver thresholds: $\eta_r \in \{0.0, 0.35\}$. In addition, the SoftMax temperature is set to $\mathcal{T} = 0.01$. To ensure our model is robust to initialisations we conducted each DoM pairing for each type of agent (e.g. DoM(-1) random sender vs. DoM(0) receiver with $\eta_r = 0.35$, DoM(-1) random sender vs. DoM(0) receiver with $\eta_r = 0.0$, etc) with 20 random seeds and then averaged the behavioural outcomes for each dyad for interpretation and visualisation. The combination of these thresholds yields asymmetric outcomes. For example, receivers with higher threshold $\eta_r = 0.35$ will act more aggressively (rejecting more offers) vs. their zero threshold counterpart. In turn, the lower threshold sender $\eta_s = 0.1$ is likely to “compromise” and improve its offers in response, vs. the higher threshold sender. Other thresholds yield similar patterns, if they satisfy $\eta_s + \eta_r < 1$. The game duration was set to allow agents sufficient time to accumulate enough evidence (likelihood) to detect the type of opponent they interact with. Previous pilot work (Alon, Schulz, Dayan, & Barnby, 2023) used shorter duration, different sender threshold, higher SoftMax temperature and no receiver threshold and yielded similar results. We present the results from this pilot work in the appendix. We also experimented with higher SoftMax temperature $\mathcal{T} = 1.0$ to evaluate the effect of high temp on detectability and consequent behaviour. These results are discussed later.

RESULTS

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

First, since higher DoM agents (such as DoM(1)) model the beliefs of lower DoM agents (such as DoM(0)) to their strategic benefit, a DoM(>1) agent can confuse naive behaviour as arising instead from a DoM(1). We account for this by analyzing the counter-deceptive reasoning applied by DoM(2) agents and show how this sophisticated strategy can be a blessing when matched with DoM(1) agents, and be disadvantageous when matched with simpler agents.

Second, DoM(2) agents believe themselves to be interacting with DoM(1) partners, and can believe that DoM(1) partners are strategically impervious to the responses of what they assume to be their DoM(0) opponents. Thus, DoM(2) agents can exhibit a form of helplessness when playing DoM(-1), when in fact they would be perfectly capable of exploiting them appropriately, using the nested model of DoM(-1) agents and planning through their behaviour. We illustrate this by simulating the DoM(2) receiver and the DoM(-1) sender. Due to the strictness of opponent reasoning of the

¹ Note that in this case the DoM levels are matched and adaptive from the perspective of the DoM($k + 1$), but this is untrue for the DoM(k) who cannot conceive of the DoM($k + 1$),s world model.

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 find that sophisticated opponents are vulnerable to over-mentalizing and subsequent loss of reward.

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 mentalize about the receiver. 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 within the bounds of any inherent irreducible uncertainty.

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

Following the properties of hierarchical mentalizing, we begin with the first dyad of adaptively aligned DoM – 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, in this example, the first offer is enough to parse whether the partner is a random sender or a threshold sender (since the offer 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, 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” 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 behaviour is appropriate given the context.

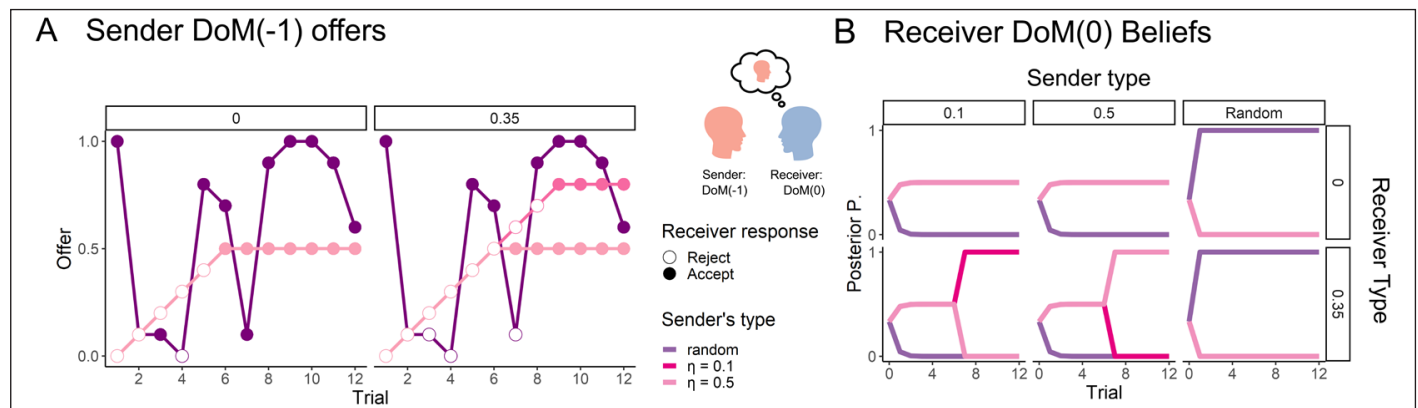


Figure 2 Illustration of DoM(0) IRL: (A, B) In interacting with the DoM(-1) sender (A), the DoM(0) receiver makes inferences about the sender's type (B). Notably, the first offer is usually sufficient to tell the random sender from the threshold senders. When the receiver's belief favours the threshold sender, the receiver manipulates the sender by rejecting the offers until a desired offer is met, according to the receiver's threshold. Both DoM(-1) threshold agents are *reactive* – that is, they respond to the behaviour of others. Hence they react similarly to the strategic behaviour of the DoM(0) until their “willingness” to be bounded is limited by their threshold (after 6 trials) – the main difference between their behaviour is the maximal offer they are willing to make. The thresholds of the agents determine the range of possible agreement – agents with higher thresholds are less willing to “compromise”. For example, agents (both receiver and sender) with higher thresholds need a more egalitarian split of the endowment compared to those with low thresholds.

Note: Posterior $P(\theta)$ means the posterior distribution of the inferring agent after observing the actions of the other agent. $P(\theta) = 0$ means that the inferring agent's belief places zero probability that the observed agents has type θ and $P(\theta) = 1$ means that the inferring agent is certain that the observed agent has type θ (when lines overlap the behaviour of the DoM(-1) sender or the updated beliefs of the DoM(0) are the same for both thresholds).

The DoM(1) sender uses its DoM(0) nested model to compute what should be an optimal policy. It emulates the DoM(0) inference process and consequently predicts the DoM(0) policy. Given the policies depicted in Figure 2, the DoM(1) sender's policy is to take actions consistent with a random DoM(-1) agent, causing the DoM(0) to accept any offer (respecting the receiver's threshold). This set of random-like actions arises through the ability of the DoM(1) to model fully the expectations, beliefs, and mentalizing capacity of the DoM(0) (Kopp, Korb, & Mills, 2018); Acting in the same way as a random sender utilizes the DoM(1)'s ability to make inferences about the DoM(0)'s IRL process. The belief manipulation allows the DoM(1) sender to avoid the strategic rejection policy of the DoM(0) receiver, as from the DoM(0) perspective, the offers are highly unlikely for a threshold sender, illustrated in Figure 3(C). Given the low SoftMax temperature, the DoM(1) correctly infers that the DoM(0) receiver would infer that any offer other than 0.0 is highly likely to come from the random sender, as the threshold DoM(-1) senders are expected to start by offering nothing.

Once false beliefs have developed in the DoM(0) (approximated also by the DoM(1)), the DoM(1) sender's policy is to repeatedly send the bare minimal offer (presented in 3(A)), to extract reward at the expense of the DoM(0). As the likelihood of a flat trajectory of offers is the same as the likelihood of any other trajectory generated by a random sender, the DoM(0) receiver is unable to tell the true random from the fake one as depicted in Figure 3(B).

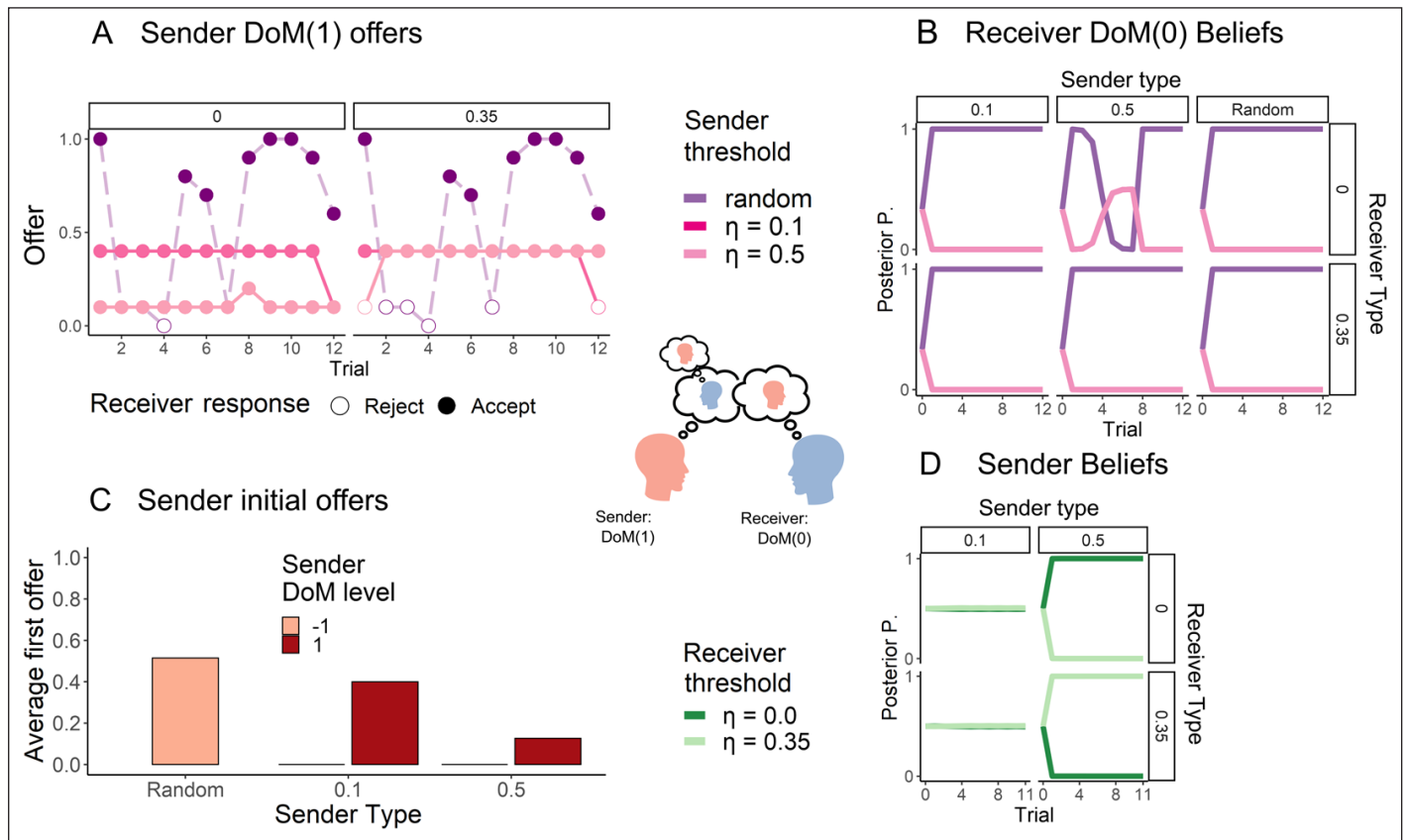


Figure 3 Dynamics of the DoM(1) manipulation: The DoM(1) offers are strategically aimed at shifting the DoM(0) belief in favour of the random sender hypothesis. This strategy naturally arises out of simple reward-maximising agents in a mixed motive setting. (A) Illustrative offers sent by different types of senders within a single seed. The sender's initial offer "mimics" that of the random sender, before subsequently defecting. (B), The DoM(0) false beliefs about the identity of the DoM(1) agent due to the strategy the DoM(1) player uses. 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.

Note: Posterior P. means the posterior distribution of the inferring agent after observing the actions of the other agent. $P = 0$ means that the inferring agent's belief places zero probability that the observed agent has type θ and $P = 1$ means that the inferring agent is certain that the observed agent has type θ (when lines overlap the behaviour of the DoM(1) sender or the updated beliefs of the DoM(0) are the same for both thresholds).

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

The DoM(2) can simulate the policy of the DoM(1), and all nested models of the DoM(1) sender. The DoM(2) can anticipate that “random” offers may arise from a sophisticated DoM(1), and thus can react accordingly (Figure 4). Applying the same belief manipulation principles as the DoM(1), the DoM(2) acts in a way that causes the DoM(1) to falsely believe that it is matched with the higher $\eta = 0.35$ DoM(0) receiver, thus pressuring the sender to improve its offers in the case of the lower threshold DoM(2) receiver. 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 (the sender has to offer at most 0.4), the DoM(2) enjoys a decrease in the sender to receiver reward ratio.

We conclude that when appropriately matched, being a DoM($k + 1$) matched with a DoM(k) partner is beneficial. These findings reinforce previous work 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.

We now examine the behavioural phenotype of a high DoM(2) receiver matched with a very simplistic DoM(-1) sender; in this case, the DoM(2) receiver is using a very sophisticated strategy for very simple sender.

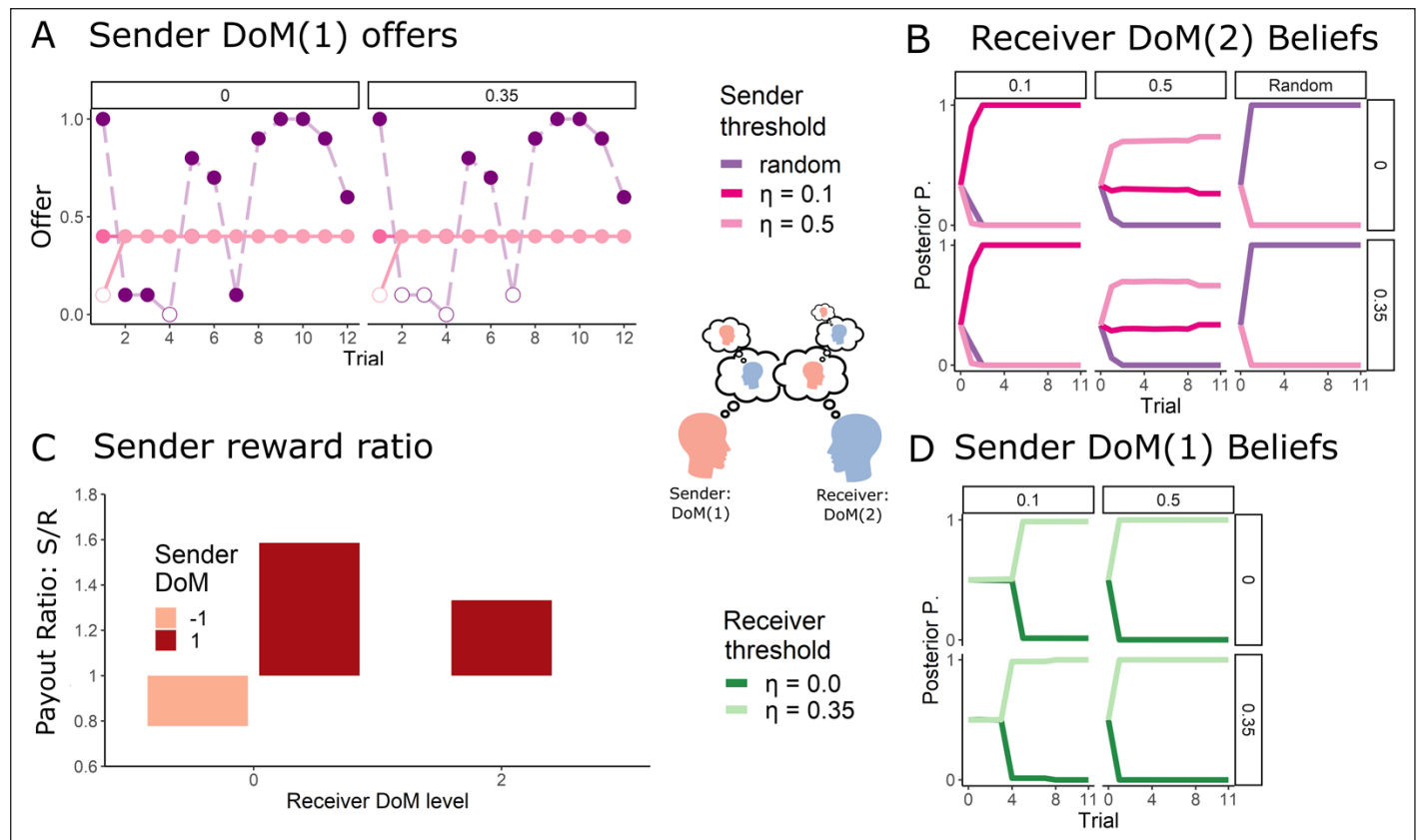


Figure 4 Dynamics of the DoM(2) counter-manipulation: (A) The DoM(2) with low threshold, masquerading as the high threshold receiver, rejects low offers. This encourages the DoM(1) sender with a high threshold to improve its offers, while having little effect on the already “generous” $\eta = 0.1$ DoM(1) sender. (B), the DoM(2) receiver correctly reads the DoM(1) sender’s strategy, while manipulating the latter’s beliefs (D) As depicted in (D), this causes the DoM(1) to develop false beliefs about the identity of the DoM(2) as they are unable to model them appropriately. (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.

Note: Posterior P. means the posterior distribution of the inferring agent after observing the actions of the other agent. $P = 0$ means that the inferring agent’s belief places zero probability that the observed agent has type θ and $P = 1$ means that the inferring agent is certain that the observed agent has type θ (when lines overlap the behaviour of the DoM(1) sender are the same for both thresholds).

As we will see, DoM is a double-edged sword. A mismatched DoM agent may misinterpret the actions of their partner, misinterpreting simplistic behaviour as the product of Machiavellian sophistication. Here, predicting that a DoM(1) sender would act like a random sender to deploy deception, the DoM(2) receiver is susceptible to interpreting random behaviour as having been generated by the DoM(1) sender, even when this is not the ground truth. This leads to delayed detection of a true random sender, as more evidence of "random" behaviour is required to confirm that the sender is genuinely random. The very possibility that random-like behaviour may be used as deception has the effect that it takes, on average, 5 trials for the DoM(2) receiver to converge to the true-random type compared to the 2 trials it takes on average for the DoM(0).

Delayed random identification has a limited effect on the DoM(2) reward when in the presence of a DoM(1). Nevertheless in the presence of a DoM(-1), the DoM(2) incurs more severe reward loss.

In cases of DoM(2) vs. DoM(-1), the over-attribution of sophisticated strategic intention to observations, typical of the DoM(1), means that the DoM(2) receiver fails to model the DoM(-1) sender. On the one hand, as mentioned, true randomness is interpreted as arising from a DoM(1). But there is also a double-bind: as the low offers of the threshold DoM(-1) senders are atypical for the random-pretending DoM(1), the DoM(2) receiver also interprets any non-random DoM(-1) actions as a sign of random behaviour as depicted in Figure 5(A).

Given that the best response for a receiver in the face of a truly random sender is to accept anything above one's threshold, the trapped DoM(2) receiver accepts most of the threshold DoM(-1) sender offers. However, the simplistic DoM(-1) 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. The DoM(2) believes one interpretation of events and thereby avoids any chance of encountering disconfirmatory evidence.

While the DoM(2) falls for the same deception applied by the DoM(1) sender against the DoM(0) receiver, the causal mechanism between the two differ. In the case of a DoM(-1) sender, the DoM(2) is a victim of its sophistication, and the incorrectly attributed sophistication of its partner, and not the victim of a truly savvy opponent.

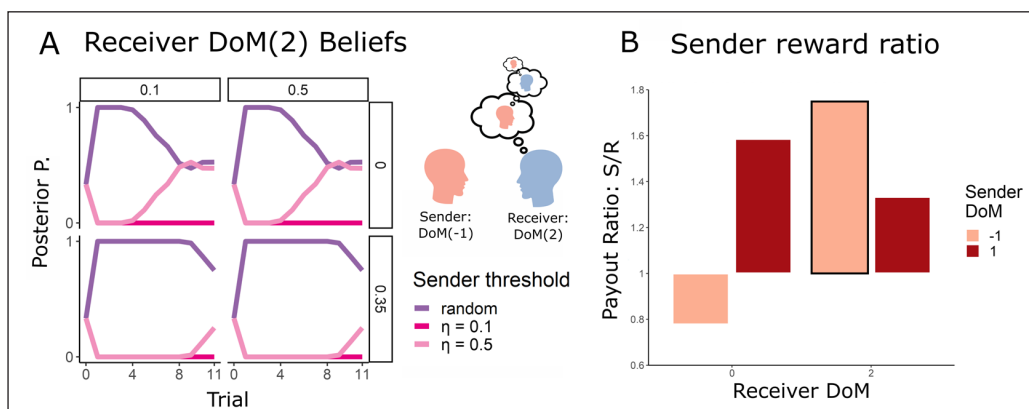


Figure 5 Effects of maladaptive DoM: (A) The DoM(2) has developed false beliefs about their unsophisticated DoM(-1) partner. This is because offers of the DoM(-1) sender lie outside the DoM(2) opponent model and are viewed as coming from a random sender. (B) In turn, the receiver's docile policy means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

Note: Posterior P. means the posterior distribution of the inferring agent after observing the actions of the other agent. $P = 0$ means that the inferring agent's belief places zero probability that the observed agent has type θ and $P = 1$ means that the inferring agent is certain that the observed agent has type θ .

INFLUENCE OF HIGH ENVIRONMENTAL NOISE

Our results highlight the attribution of *intention* to random behaviour in sophisticated agents. The DoM(2) maps the random-like behaviour of the DoM(1) as an intention to deceive. At the same time, the DoM(0), lacking the capacity to simulate such an opponent, attributes the same behaviour to a random sender. However, this detectability of intention is plausible due to the sender's close mapping of their beliefs to their actions – we chose to keep the decision temperature low. This means that from the DoM(0)'s perspective, the DoM(-1) threshold senders' behaviour is quite predictable, hence easy to invert. The question arises: Can agents still infer intentions if the environment is stochastic and noisy? To foreshadow, we show that noise is necessary but insufficient to induce the same false beliefs of malice in the DoM(2) receiver.

To examine this issue, we simulated the dyads with a high SoftMax temperature $\mathcal{T} = 1.0$. In this case, the non-random agents' behaviour is more stochastic, making Bayesian IRL hard as the mapping between beliefs and actions is tenuous. We begin with the DoM(0) baseline. In the high-temperature environment, the receiver is barely able to detect the correct type of the sender, as evident in Figure 6 (Top row). When interacting with the DoM(1), the DoM(0) is again still “fooled by randomness”, inferring the DoM(1) as a random sender, but with a lower degree of certainty, as depicted in Figure 6 (Bottom row).

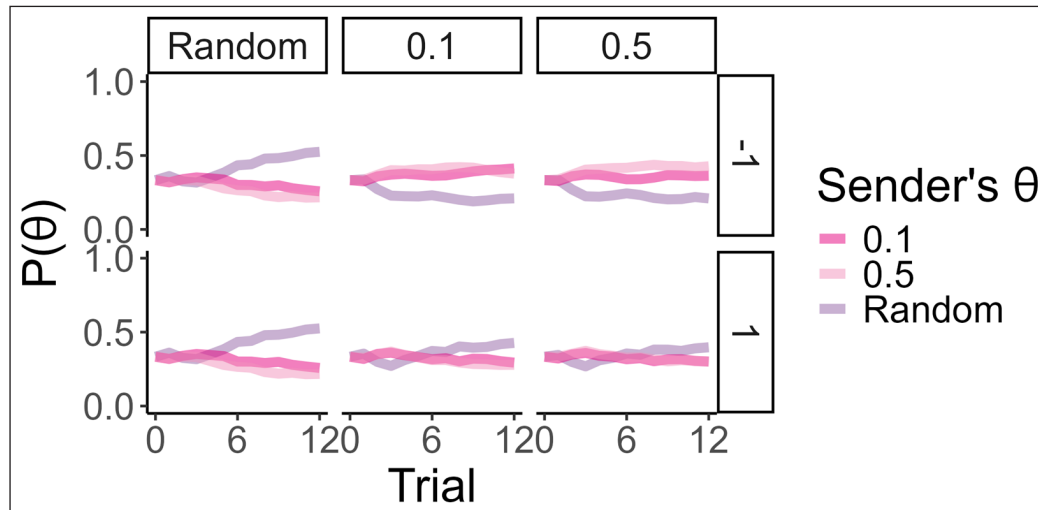


Figure 6 DoM(0) belief update in high SoftMax temperature environment. We depict the updated DoM(0) beliefs against senders with different DoM level (row indicate sender's DoM level, column indicate sender's type) averaged across 20 different simulations. Due to the noisy behaviour of the senders, the DoM(0) finds it hard to identify the sender's correct beliefs from its actions, even when it interacts with an adaptively matched sender (DoM(-1), top row). When interacting with the higher DoM sender, the receiver is still deceived, but with a lower certainty.

The DoM(2) inference yields different results, presented in Figure 7. To surmise, DoM(2) receivers view the stochastic actions as unintentional. By inverting the offers of a random sender through the lens of the DoM(1), the DoM(2) belief update takes into account the nested DoM(0) beliefs, and thus the poor identification of the sender. Consequently, the DoM(2) falsely attributes the stochastic behaviour to a noisy DoM(1) threshold sender as evident in 7. This misidentification is caused by the DoM(2) receiver inferring offers as stemming from a low-threshold sender – the offers are “too high” to be counted as being generated by a high-threshold sender, even in a stochastic environment.

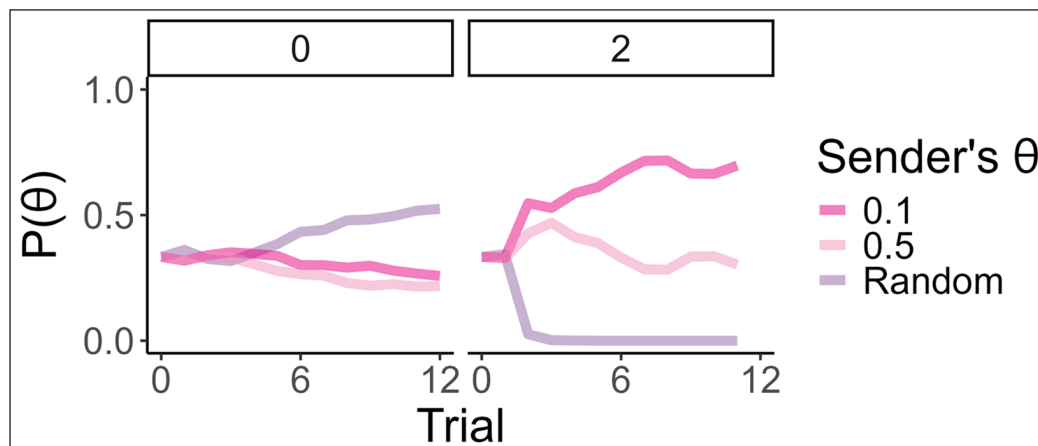


Figure 7 Comparison of updated beliefs against a random DoM(1) sender in high SoftMax environment. We depict the updated receiver beliefs interacting with a random sender averaged across 20 different simulations. The DoM(0) attributes the random behaviour of the random sender (left panel). On the other hand, the DoM(2) receiver attributes the random behaviour to a benign, high-temperature DoM(1) sender.

DISCUSSION

We show that hierarchical mentalizing is a double-edged sword. We analysed pairs of RL agents endowed with ToM at different depths of mentalizing 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

(Lyons, Caldwell, & Shultz, 2010; Whiten & Byrne, 1988; Qi & Vul, 2020). 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 an over- or hyper-mentalized policy, believing they are matched with a sophisticated other that will try to deceive. Importantly, we show that both ambiguous behaviour and intentionality are necessary and sufficient to explain maladaptive overmentalizing – the former in the absence of the latter is explained away as unintentional noise. 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 is a product of the agent's own beliefs, its environmental context, and the internal representation of an other. This is consistent with prior observations (Simon, 1990; Bhui, Lai, & Gershman, 2021; Huys, Guitart-Masip, Dolan, & Dayan, 2015), and is relevant for the maladaptive behaviour of machines (E. Schulz & Dayan, 2020). It also shows how complex phenomena like scepticism can arise even from optimal Bayesian inference (Bhui & Gershman, 2020; Alon, Schulz, Rosenschein, & Dayan, 2023) and how optimal Bayesian inference can go awry given confusion about a decision problem or an unfortunate environment (Huys, Guitart-Masip, Dolan, & Dayan, 2015).

The over-attribution of negative social intentions is a central feature in paranoid delusions and borderline personality disorder (Buck et al., 2023) and indeed hyper-mentalizing has been identified as an important transdiagnostic feature in psychopathology more broadly (McLaren et al., 2022; Sharp et al., 2011). Our work offers a computational model of these phenomena, formalising a theory of how hierarchical, recursive social cognition gone awry may explain the emergence of paranoia, which may be maintained in purely reward-maximising, interactive agents with minor miscalibrations. This cognitive mechanism may play a crucial role in the formation of persecutory delusions, along with inflexible priors about interaction partners (Barnby, Raihani, & Dayan, 2022; Diaconescu, Wellstein, Kasper, Mathys, & Stephan, 2020; Wellstein et al., 2020), noisy mental models of others (Adams, Vincent, Benrimoh, Friston, & Parr, 2022; Barnby, Mehta, & Moutoussis, 2022), social hypersensitivity (Henco et al., 2020), and biased social values (Kazinka, Kwashie, Pratt, Vilares, & MacDonald III, 2023). Of note, our models also offer a glimpse into the emergence of biases against disconfirmatory evidence (Freeman, 2016), given that hypermentalized receivers in our simulations form their fixed beliefs quickly, and thus fail to expose themselves to actions which may cause belief reorganisation.

Crucially, the DoM(2) discerns between intended and unintended decision noise. The DoM(2) infers stochasticity as a sign of intended strategy only in the low SoftMax environment – it attributed random behaviour as an intentional choice by the strategic DoM(1) partner. In contrast, in the high-temp environment, the DoM(2) infers the stochastic behaviour as an unintentional noisy execution of the DoM(1)'s policy. Our model explains why noisy environmental expectations have typically been associated with persecutory beliefs in reward tasks (e.g. Hauke et al., 2024) and in traditional vignette tasks (e.g. Kinderman, Dunbar, & Bentall, 1998), but highlights that this in of itself is insufficient to explain the intentionality of attributions of intentional harm. The addition of our high-SoftMax simulations shows a distinction: 1. noise is necessary, and 2. it must be viewed as intentional – a characteristic only possible in DoM(2) agents.

The results of the simulations pave the path for human experiments. Paradigms testing this theory directly should use the IUG in their designs, with computerized partners (senders) calibrated across differing levels of DoM from $k = -1$ to $k = 2$. Participants will play different partners in a within-subject design over multiple games, each game consisting of at least 12 trials as per our simulations. Given parallel work in social interaction and hierarchical cognition (Bürgi, Aydogan, Kononov, & Ruff, 2024; Qi & Vul, 2020) we would expect most participants to be successful in adapting their DoM to the complexity of their partner. Nevertheless, we would expect some deviations: 1. High trait paranoia will be significantly associated with mismatched DoM, such that simple partners (DoM($k = -1$) or DoM($k = 0$)) are met with relatively high DoM (DoM($k = 1$) or DoM($k = 2$), respectively) by the participant (ascertained through model fitting). Testing this paradigm in

patients with persecutory delusions will be a core aspiration. We would expect that patients with persecutory delusions will make the same errors as those with high trait paranoia in the general population (demonstrating overly deep DoM), although do so with greater frequency and rigidity (taking longer, if at all, to match their partner's DoM appropriately). We would also expect D2/D3 antagonism to reduce this tendency in both general (Barnby, Bell et al., 2023) and clinical (Adams et al., 2022) populations. Across the board, patients with persecutory delusions and those with high trait paranoia should be equally effective as other participants when matched with senders at DoM(1) who are sophisticated and deceptive. In this case, occupying a higher DoM is adaptive.

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) (Fehr & Schmidt, 1999) or FS-like utility functions are typically used to assess rejection in social contexts (e.g. Hula et al., 2018; Kazinka et al., 2023), although we opted to remove this to isolate the effect of DoM. Replacing these egocentric utilities with social orientation utilities, like inequity aversion (Hughes et al., 2018), may yield other non-trivial effects of hypermentalizing.

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-mentalization 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 (Camerer et al., 2004), 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-mentalization, as the higher DoM agent no longer treats others as having a fixed DoM($k - 1$) but rather has multiple (unknown) DoM levels. However, this instantiation comes with an increase in the computational costs and complexity of the inference process.

One future direction for solving fixed over-mentalization 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 delay in making this reduction even when the computational and utility costs are high. 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 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 be unable 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 mentalization 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 before drastic action, e.g., gathering more information about opponents (L. Schulz, Fleming, & Dayan, 2023).

APPENDIX

EXPERIMENTS WITH OTHER PARAMETERS

In (Alon, Schulz, Dayan, & Barnby, 2023), the IUG task was simulated with different set of parameters $\eta_s \in \{0.1, 0.5\}$, $\eta_r = 0.0$, $\mathcal{T} = 0.05$ and $T = 10$. The main findings are similar to the ones presented in this work and are displayed in Figure 8.

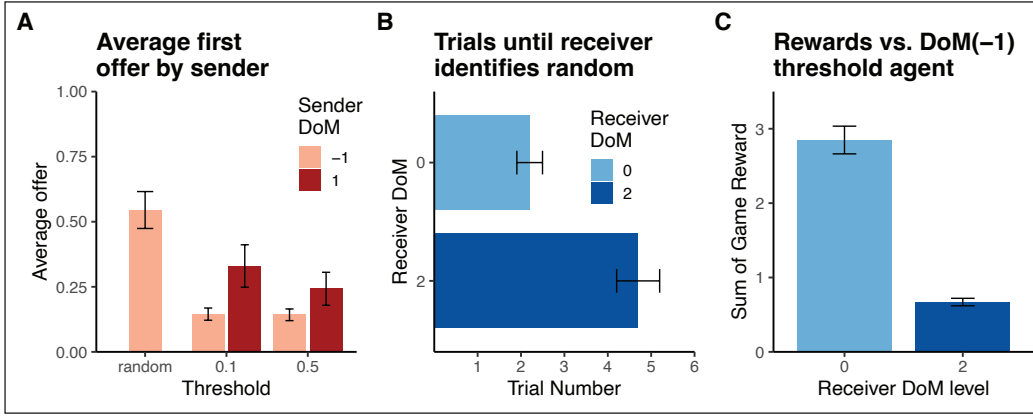


Figure 8 Results summary:

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, 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.

IRL AS TOM INFERENCE

As presented above, inference in ToM can be seen as an extension of inverse RL (IRL) (Ng & Russell, 2000). Bayesian IRL (Ramachandran & Amir, 2007; C. Baker et al., 2011) requires an observer to make a Bayesian inference about the utility (reward) function of an agent from a sequence of observed behaviour $\sigma^{0:T}$:

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

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 (Jara-Ettinger, 2019), and framed as the Naïve utility calculus (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). 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 Figure 2B.

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 (Hjort, Holmes, Müller, & Walker, 2010). We refer the reader to (Gmytrasiewicz & Doshi, 2005) for a full introduction of belief update in this case.

FUNDING INFORMATION

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 “ClinbrAIIn: Artificial Intelligence for Clinical Brain Research”. JMB has received funding from the Wellcome Trust (WT228268/Z/23/Z) and is supported as a FENS-Kavli Scholar as part of the FENS-Kavli Network of Excellence.

COMPETING INTERESTS

The authors have no competing interests to declare.

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.

AUTHOR AFFILIATIONS

Nitay Alon  orcid.org/0000-0002-2698-3573

Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel; Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Lion Schulz  orcid.org/0000-0003-1841-1273

Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Vaughan Bell  orcid.org/0000-0001-8616-4847

Clinical, Educational, and Health Psychology, University College London, United Kingdom

Michael Moutoussis  orcid.org/0000-0002-4751-0425

Department of Imaging Neuroscience, University College London, London, United Kingdom

Peter Dayan  orcid.org/0000-0003-3476-1839

Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany; Department of Computer Science, University of Tübingen, Tübingen, Germany

Joseph M. Barnby  orcid.org/0000-0001-6002-1362

Department of Psychology, Royal Holloway University of London, London, United Kingdom; School of Psychiatry and Clinical Neuroscience, The University of Western Australia, Australia

REFERENCES

- Adams, R. A., Vincent, P., Benrimoh, D., Friston, K. J., & Parr, T. (2022). Everything is connected: inference and attractors in delusions. *Schizophrenia research*, 245, 5–22. <https://doi.org/10.1016/j.schres.2021.07.032>
- Alon, N., Schulz, L., Dayan, P., & Barnby, J. M. (2023). Between prudence and paranoia: Theory of mind gone right, and wrong. In *First workshop on theory of mind in communicating agents*. <https://openreview.net/forum?id=gB9zrEjhZD>
- Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023, 08). A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind*, 7, 608–624. https://doi.org/10.1162/opmi_a_00097
- Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33). Retrieved 2022-08-23 from <https://escholarship.org/uc/item/5rk7z59q>
- Baker, C., & Tenenbaum, J. (2014, March). Modeling Human Plan Recognition Using Bayesian Theory of Mind. *Plan, Activity, and Intent Recognition: Theory and Practice*, 177–204. <https://doi.org/10.1016/B978-0-12-398532-3.00007-5>
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, March). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 1–10. Retrieved 2020-05-28 from <https://www.nature.com/articles/s41562-017-0064> (Number: 4 Nature Publishing Group).
- Barnby, J. M., Bell, V., Deeley, Q., Mehta, M., & Moutoussis, M. (2023). D2/d3 dopamine supports the precision of mental state inferences and self-relevance of joint social outcomes. *bioRxiv*, 2023–05.
- Barnby, J. M., Bellucci, G., Alon, N., Schilbach, L., Bell, V., Frith, C., & Dayan, P. (2023). Beyond theory of mind: A formal framework for social inference and representation. *PsyArXiv*. <https://doi.org/10.31234/osf.io/cmgu7>
- Barnby, J. M., Mehta, M. A., & Moutoussis, M. (2022). The computational relationship between reinforcement learning, social inference, and paranoia. *PLoS computational biology*, 18(7), e1010326. <https://doi.org/10.1371/journal.pcbi.1010326>
- Barnby, J. M., Raihani, N., & Dayan, P. (2022). Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, 225, 105098. <https://doi.org/10.1016/j.cognition.2022.105098>
- Barnett, S. A., Griffiths, T. L., & Hawkins, R. D. (2022). A pragmatic account of the weak evidence effect. *Open Mind*, 1–14. https://doi.org/10.1162/opmi_a_00061
- Bell, V., Mills, K. L., Modinos, G., & Wilkinson, S. (2017). Rethinking social cognition in light of psychosis: reciprocal implications for cognition and psychopathology. *Clinical Psychological Science*, 5(3), 537–550. <https://doi.org/10.1177/2167702616677079>
- Bentall, R., & Kinderman, P. (1998). Psychological processes and delusional beliefs: Implications for the treatment of paranoid states. *Outcome and innovation in psychological treatment of schizophrenia*.
- Bhui, R., & Gershman, S. J. (2020). Paradoxical effects of persuasive messages. *Decision*, 7(4), 239–258. <https://doi.org/10.1037/dec0000123>

- Bhui, R., Lai, L., & Gershman, S. J. (2021, October). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41, 15–21. Retrieved 2023-04-21, from <https://www.sciencedirect.com/science/article/pii/S2352154621000371>
- Bowes, S. M., Costello, T. H., & Tasimi, A. (2023). The conspiratorial mind: A meta-analytic review of motivational and personological correlates. *Psychological Bulletin*. <https://doi.org/10.1037/bul0000392>
- Buck, B., Browne, J., Gagen, E. C., & Penn, D. L. (2023). Hostile attribution bias in schizophrenia-spectrum disorders: narrative review of the literature and persisting questions. *Journal of Mental Health*, 32(1), 132–149. <https://doi.org/10.1080/09638237.2020.1739240>
- Bürgi, N., Aydogan, G., Konovalov, A., & Ruff, C. C. (2024, Feb). A neural fingerprint of adaptive mentalization. *PsyArXiv*. osf.io/preprints/psyarxiv/npqy8 <https://doi.org/10.31234/osf.io/npqy8>
- Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3), 861–898. <https://doi.org/10.1162/0033553041502225>
- Chan, K. K., & Chen, E. Y. (2011). Theory of mind and paranoia in schizophrenia: a game theoretical investigation framework. *Cognitive Neuropsychiatry*, 16(6), 505–529. <https://doi.org/10.1080/13546805.2011.561576>
- Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., & Montague, P. (2008). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, 57(3), 463–473. <https://www.sciencedirect.com/science/article/pii/S0896627307010331>
- Corcoran, R., Rowse, G., Moore, R., Blackwood, N., Kinderman, P., Howard, R., ... Bentall, R. (2007). A transdiagnostic investigation of ‘theory of mind’ and ‘jumping to conclusions’ in patients with persecutory delusions. *Psychological Medicine*, 38, 1577–1583. <https://doi.org/10.1017/S0033291707002152>
- Dennett, D. C. (1989). *The intentional stance*. MIT press. <https://doi.org/10.1017/S0140525X00058611>
- Devaine, M., Hollard, G., & Daunizeau, J. (2014a, December). Bayesian Brain: Does Mentalizing Make a Difference When We Learn? The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? *PLOS Computational Biology*, 10(12), e1003992. Retrieved 2023-01-25 from <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003992> (Public Library of Science)
- Devaine, M., Hollard, G., & Daunizeau, J. (2014b, 02). Theory of mind: Did evolution fool us? *Plos One*, 9(2), 1–12. <https://doi.org/10.1371/journal.pone.0087619>
- Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., & Stephan, K. E. (2020). Hierarchical bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, 129(6), 556. <https://doi.org/10.1037/abn0000500>
- Doshi, P., Qu, X., & Goodie, A. (2014, January). Chapter 8- Decision-Theoretic Planning in Multiagent Settings with Application to Behavioral Modeling. In G. Sukthankar, C. Geib, H. H. Bui, D. V. Pynadath, & R. P. Goldman (Eds.), *Plan, Activity, and Intent Recognition* (pp. 205–224). Boston Morgan Kaufmann. Retrieved 2023-02-27 from <https://www.sciencedirect.com/science/article/pii/B9780123985323000087>
- Euler, S., Nolte, T., Constantinou, M., Griem, J., Montague, P. R., Fonagy, P., ... Network, M. D. R. (2021). Interpersonal problems in borderline personality disorder: associations with mentalizing, emotion regulation, and impulsiveness. *Journal of Personality Disorders*, 35(2), 177–193. https://doi.org/10.1521/pedi_2019_33_427
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868. Retrieved 2023-06-22 from <http://www.jstor.org/stable/2586885>
- Fett, A.-K. J., Shergill, S. S., Joyce, D. W., Riedl, A., Strobel, M., Gromann, P. M., & Krabbendam, L. (2012). To trust or not to trust: the dynamics of social interaction in psychosis. *Brain*, 135(3), 976–984. <https://doi.org/10.1093/brain/awr359>
- Fett, A.-K. J., Viechtbauer, W., Penn, D. L., van Os, J., Krabbendam, L., et al. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. *Neuroscience & Biobehavioral Reviews*, 35(3), 573–588. <https://doi.org/10.1016/j.neubiorev.2010.07.001>
- Fonagy, P., Luyten, P., & Bateman, A. (2015). Translation: Mentalizing as treatment target in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, 6(4), 380. <https://doi.org/10.1037/per0000113>
- Freeman, D. (2016). Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7), 685–692. [https://doi.org/10.1016/S2215-0366\(16\)00066-3](https://doi.org/10.1016/S2215-0366(16)00066-3)
- Frith, U., & Happé, F. (1994). Autism: beyond “theory of mind”. *Cognition*, 50(1), 115–132. <https://www.sciencedirect.com/science/article/pii/0010027794900248>
- Galvez-Merlin, A., Lopez-Villatoro, J. M., de la Higuera-Gonzalez, P., de la Torre-Luque, A., Reneses-Prieto, B., Diaz-Marsa, M., & Carrasco, J. L. (2023). Social cognition deficits in borderline personality disorder: Clinical relevance. *Psychiatry Research*, 115675. <https://doi.org/10.1016/j.psychres.2023.115675>

- Gmytrasiewicz, P. J., & Doshi, P.** (2005, July). A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24, 49–79. Retrieved 2020-03-10 from <https://jair.org/index.php/jair/article/view/10414>
- Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. B.** (2006). 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).
- Goodman, N. D., & Frank, M. C.** (2016, November). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11), 818–829. Retrieved 2023-09-26 from <https://www.sciencedirect.com/science/article/pii/S136466131630122X>
- Hauke, D. J., Wobmann, M., Andreou, C., Mackintosh, A. J., de Bock, R., Karvelis, P., ... others** (2024). Altered perception of environmental volatility during social learning in emerging psychosis. *Computational Psychiatry*, 8(1), 1. <https://doi.org/10.5334/cpsy.95>
- Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M.-L., Hörmann, S., Hennings, J., ... others** (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS computational biology*, 16(9), e1008162. <https://doi.org/10.1371/journal.pcbi.1008162>
- Hertwig, R., & Engel, C.** (2016). Homo Ignorans: Deliberately Choosing Not to Know. *Perspectives on Psychological Science*, 11(3), 359–372. <https://doi.org/10.1177/1745691616635594>
- Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G.** (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge University Press. <https://doi.org/10.1017/CBO9780511802478>
- Ho, M. K., Saxe, R., & Cushman, F.** (2022, September). Planning with Theory of Mind. *Trends in Cognitive Sciences*. Retrieved 2022-09-09 from <https://www.sciencedirect.com/science/article/pii/S1364661322001851>
- Hughes, E., Leibo, J. Z., Phillips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., ... Graepel, T.** (2018, September). Inequity aversion improves cooperation in intertemporal social dilemmas. [*cs, q-bio*]. Retrieved 2021-03-07 from <http://arxiv.org/abs/1803.08884> (arXiv: 1803.08884)
- Hula, A., Montague, P. R., & Dayan, P.** (2015a). Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange. *PLoS Computational Biology*, 11(6), e1004254. <https://doi.org/10.1371/journal.pcbi.1004254>
- Hula, A., Montague, P. R., & Dayan, P.** (2015b, June). Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange. *PLOS Computational Biology*, 11(6), e1004254. Retrieved 2023-03-23 from <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004254> (Public Library of Science) <https://doi.org/10.1371/journal.pcbi.1004254>
- Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R.** (2018, 02). A model of risk and mental state shifts during social interaction. *PLOS Computational Biology*, 14(2), 1–20. <https://doi.org/10.1371/journal.pcbi.1005935>
- Hutter, M.** (2004). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.
- Huys, Q. J., Guitart-Masip, M., Dolan, R. J., & Dayan, P.** (2015, May). Decision-Theoretic Psychiatry. *Clinical Psychological Science*, 3(3), 400–421. (SAGE Publications Inc.) <https://doi.org/10.1177/2167702614562040>
- Jara-Ettinger, J.** (2019, October). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29, 105–110. Retrieved 2023-03-15 from <https://www.sciencedirect.com/science/article/pii/S2352154618302055>
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B.** (2016, August). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, 20(8), 589–604. Retrieved 2022-03-15 from <https://linkinghub.elsevier.com/retrieve/pii/S1364661316300535>
- Kazinka, R., Kwashie, A. N., Pratt, D., Vilares, I., & MacDonald, A. W. III** (2023). Value representations of spite sensitivity in psychosis on the minnesota trust game. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. <https://doi.org/10.1016/j.bpsc.2023.11.010>
- Kinderman, P., Dunbar, R., & Bentall, R. P.** (1998). Theory-of-mind deficits and causal attributions. *British journal of Psychology*, 89(2), 191–204. <https://doi.org/10.1111/j.2044-8295.1998.tb02680.x>
- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R.** (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806–810. <https://doi.org/10.1126/science.1156902>
- Kopp, C., Korb, K. B., & Mills, B. I.** (2018, November). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to “fake news”. *Plos One*, 13(11), e0207383. (Public Library of Science) Retrieved 2022-08-08 from <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207383>

- Lyons, M., Caldwell, T., & Shultz, S. (2010). Mind-reading and manipulation—is machiavellianism related to theory of mind? *Journal of Evolutionary Psychology*, 8(3), 261–274. <https://doi.org/10.1556/JEP.8.2010.3.7>
- McLaren, V., Gallagher, M., Hopwood, C. J., & Sharp, C. (2022). Hypermentalizing and borderline personality disorder: A meta-analytic review. *American Journal of Psychotherapy*, 75(1), 21–31. <https://doi.org/10.1176/appi.psychotherapy.20210018>
- Ng, A. Y., & Russell, S. (2000). Algorithms for Inverse Reinforcement Learning. In *Proc. 17th International Conf. on Machine Learning* (pp. 663–670). Morgan Kaufmann.
- O’Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, 36(4), 313–322. <https://doi.org/10.1016/j.evolhumbehav.2015.01.004>
- Penn, D. L., Sanna, L. J., & Roberts, D. L. (2008). Social cognition in schizophrenia: an overview. *Schizophrenia bulletin*, 34(3), 408–411. <https://doi.org/10.1093/schbul/sbn014>
- Premack, D., & Woodruff, G. (1978, December). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. (Cambridge University Press). Retrieved 2020-08-10 from <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0>
- Qi, W., & Vul, E. (2020). Adaptive behavior in variable games requires theory of mind. *PsyArXiv*. <https://doi.org/10.31234/osf.io/7kw4z>
- Ramachandran, D., & Amir, E. (2007, January). Bayesian inverse reinforcement learning. In *Proceedings of the 20th international joint conference on Artificial intelligence* (pp. 2586–2591). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Randall, F., Corcoran, R., Day, J., & Bentall, R. (2003). Attention, theory of mind, and causal attributions in people with persecutory delusions: A preliminary investigation. *Cognitive neuropsychiatry*, 8(4), 287–294. <https://doi.org/10.1080/1354680000057>
- Ray, D., King-Casas, B., Montague, P., & Dayan, P. (2008). Bayesian model of behaviour in economic games. *Advances in neural information processing systems*, 21.
- Rifkin-Zybutz, R., Moran, P., Nolte, T., Feigenbaum, J., King-Casas, B., Personality, L., ... Montague, R. (2021). Impaired mentalizing in depression and the effects of borderline personality disorder on this relationship. *Borderline personality disorder and emotion dysregulation*, 8(1), 15. <https://doi.org/10.1186/s40479-021-00153-x>
- Schulz, E., & Dayan, P. (2020). Computational Psychiatry for Computers. *iScience*, 23(12), 101772. <https://doi.org/10.1016/j.isci.2020.101772>
- Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive Computations for Information Search: Confidence in Control. *Psychological Review*. <https://doi.org/10.1037/rev0000401>
- Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). 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. <https://doi.org/10.1016/j.jaac.2011.01.017>
- Simon, H. A. (1990). Invariants of Human Behavior. *Annual review of psychology*, 41(1). <https://doi.org/10.1146/annurev.ps.41.020190.000245>
- Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E. A., ... Stephan, K. E. (2020). Inflexible social inference in individuals with subclinical persecutory delusional tendencies. *Schizophrenia Research*, 215, 344–351. <https://doi.org/10.1016/j.schres.2019.08.031>
- Whiten, A., & Byrne, R. W. (1988). The machiavellian intelligence hypotheses.
- Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, 33(3), 1099–1108. <https://doi.org/10.1523/JNEUROSCI.1642-12.2013>
- Yoshida, W., Dziobek, I., Kliemann, D., Heekeren, H. R., Friston, K. J., & Dolan, R. J. (2010). Cooperation and heterogeneity of the autistic mind. *Journal of Neuroscience*, 30(26), 8815–8818. <https://doi.org/10.1523/JNEUROSCI.0400-10.2010>

TO CITE THIS ARTICLE:

Alon, N., Schulz, L., Bell, V., Moutoussis, M., Dayan, P., & Barnby, J. M. (2024). (Mal) adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1), pp. 159–177. DOI: <https://doi.org/10.5334/cpsy.117>

Submitted: 20 March 2024

Accepted: 29 August 2024

Published: 11 September 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Computational Psychiatry is a peer-reviewed open access journal published by Ubiquity Press.

Chapter 4

Paper III: $\aleph$ -IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection

Published as: Alon, N., Barnby, J. M., Sarkadi, S., Schulz, L., Rosenschein, J. S., & Dayan, P. (2026). $\aleph$ -IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection. *Journal of Artificial Intelligence Research*, Vol. 85, Article 14. <https://doi.org/10.1613/jair.1.19204>

Author contributions: N.A. is the corresponding author and led the conceptualisation, formal framework, implementation, and writing. All co-authors contributed to conceptualisation and writing.

N-IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection

NITAY ALON*, Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Germany, School of Computer Science and Engineering The Hebrew University of Jerusalem, Israel

JOSEPH M. BARNBY, Centre for AI and Machine Learning Edith Cowan University, Australia, Institute of Psychiatry, Psychology and Neuroscience King's College London, UK

STEFAN SARKADI, Centre for Defence and Security Artificial Intelligence University of Lincoln, UK

LION SCHULZ, Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Germany

JEFFREY S. ROSENSCHEIN, School of Computer Science and Engineering The Hebrew University of Jerusalem, Israel

PETER DAYAN, Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Germany, Department of Computer Science University of Tübingen, Germany

Social agents with finitely nested opponent models are vulnerable to manipulation by agents with deeper recursive capabilities. This imbalance, rooted in logic and the theory of recursive modelling frameworks, cannot be solved directly. We propose a computational framework called N-IPOMDP, which augments the Bayesian inference of model-based RL agents with an anomaly detection algorithm and an out-of-belief policy. Our mechanism allows agents to realize that they are being deceived, even if they cannot understand how, and to deter opponents via a credible threat. We test this framework in both a mixed-motive and a zero-sum game. Our results demonstrate the N-mechanism's effectiveness, leading to more equitable outcomes and less exploitation by more sophisticated agents. We discuss implications for AI safety, cybersecurity, cognitive science, and psychiatry.

JAIR Associate Editor: Davide Grossi

JAIR Reference Format:

Nitay Alon, Joseph M. Barnby, Stefan Sarkadi, Lion Schulz, Jeffrey S. Rosenschein, and Peter Dayan. 2026. N-IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection. *Journal of Artificial Intelligence Research* 85, Article 14 (February 2026), 33 pages. doi: [10.1613/jair.1.19204](https://doi.org/10.1613/jair.1.19204)

*Corresponding Author.

Authors' Contact Information: Nitay Alon, ORCID: [0000-0002-2698-3573](https://orcid.org/0000-0002-2698-3573), nitay.alon@mail.huji.ac.il, Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Tübingen, Germany, and School of Computer Science and Engineering The Hebrew University of Jerusalem, Jerusalem, Israel; Joseph M. Barnby, ORCID: [0000-0001-6002-1362](https://orcid.org/0000-0001-6002-1362), Centre for AI and Machine Learning Edith Cowan University, Perth, Australia, and Institute of Psychiatry, Psychology and Neuroscience King's College London, London, UK; Stefan Sarkadi, ORCID: [0000-0003-3999-528X](https://orcid.org/0000-0003-3999-528X), Centre for Defence and Security Artificial Intelligence University of Lincoln, Lincoln, UK; Lion Schulz, ORCID: [0000-0003-1841-1273](https://orcid.org/0000-0003-1841-1273), Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Tübingen, Germany; Jeffrey S. Rosenschein, ORCID: [0000-0002-4042-9739](https://orcid.org/0000-0002-4042-9739), School of Computer Science and Engineering The Hebrew University of Jerusalem, Jerusalem, Israel; Peter Dayan, ORCID: [0000-0003-3476-1839](https://orcid.org/0000-0003-3476-1839), Department of Computational Neuroscience Max Planck Institute for Biological Cybernetics, Tübingen, Germany, and Department of Computer Science University of Tübingen, Tübingen, Germany.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2026 Copyright held by the owner/author(s).

doi: [10.1613/jair.1.19204](https://doi.org/10.1613/jair.1.19204)

1 Introduction

Deception is a constant in human and animal cultures. Humans use a range of deceptive techniques, from “white lies” to malicious and harmful manipulation, misdirecting the beliefs of others for their benefit. To manipulate, a deceiver needs to both create false beliefs and avoid disclosing the deceiver’s true intentions. Agents can achieve this through perspective taking, a form of Theory of Mind (ToM; [Premack and Woodruff 1978](#)).

ToM encompasses the capacity to simulate others’ actions and beliefs. This can be shallow, e.g., observational learning, or recursive (i.e., including the other’s capacity to simulate the self, and so forth). The degree to which an agent can use recursive beliefs is known as its depth of mentalising (DoM; [Barnby et al. 2023](#); [C. D. Frith and U. Frith 2021](#)). This property has been explained through formal models. A successful and popular framing is k-level hierarchical ToM ([Camerer et al. 2004](#)). K-level ToM predicts that agents with lower DoM are formally incapable of making accurate inferences about the intentions of those with higher DoM ([Gmytrasiewicz and Doshi 2005](#)). Such an ability would suggest that agents had circumvented the paradox of self-reference. This limitation, found in all recursive modelling frameworks ([Pacuit and Roy 2017](#)), implies that agents with low DoM are doomed to be manipulated by others with higher DoM. This asymmetry has previously been explored ([Alon, Schulz, Dayan, et al. 2023](#); [Hula, Vilares, et al. 2018](#); [Ş. Sarkadi, Panisson, et al. 2019](#); [Ş. Sarkadi, Rutherford, et al. 2021](#); [Weerd et al. 2022](#)), illustrating the various ways that higher DoM agents can take advantage of lower DoM interaction partners.

While these agents are necessarily disadvantaged, all is not lost. Low DoM agents may still notice that the behaviour they *observe* is inconsistent with the behaviour they *expect*, even if they lack the knowledge to understand how or why ([Hula, Vilares, et al. 2018](#)). This type of mismatch warns the victim that they are facing an unmodeled opponent, meaning they can no longer use their simulation of an other for optimal planning.

One way to combat this apparent model failure is to switch to an out-of-belief (OOB) policy, where actions are detached from inferences about their opponent. One path from this is to switch from exploit to exploration behaviour, opening the way to the development of a new opponent model. However, if losing carries a detrimental impact, another option may be to take defensive action, choosing to avoid or reject an environment.

In this work, we present a computational framework for multi-agent RL (MARL) called **Σ**-IPOMDP. We augment the well-known IPOMDP approach ([Gmytrasiewicz and Doshi 2005](#)) to allow agents to engage with unmodeled opponents. We first discuss how agents can use higher DoM to manipulate others, preying on the limited modelling capacity of their partners. We then present the main contribution of this work: a deception detection mechanism, the **Σ**-mechanism, by which limited agents with shallow DoM use an off-policy counterfactual mechanism to detect anomalous behaviour, indicating that they are being deceived, and the OOB **Σ**-policy, aimed at reasonably responding to the unknown opponent. We illustrate this mechanism in two game environments, mixed-motive and zero-sum Bayesian repeated games, to show how **Σ**-IPOMDP agents can mitigate the advantages of agents with deeper recursive models, and successfully deter manipulation.

Our work is relevant to multiple fields. To the MARL community, we show how agents with limited opponent modelling (for example, suffering from bounded rationality) can partially cope with more adequate opponents via anomaly detection and game-theoretic principles. To the cybersecurity community, we present a MARL masquerading detection use case which can be used to overcome learning adversarial attacks ([Rosenberg et al. 2021](#)). To the psychology community, we provide an insights into how humans may use heuristics to detect deception when high DoM resources are not available, and why humans can avoid deception even in the absence of complex recursive reasoning. Given this framework, we can also show how human anomaly detection mechanisms may become maladaptively sensitive and overestimate deception in social environments without the need for over-mentalising, providing a context-appropriate model of suspicious or conspiratorial thinking. Lastly, there has recently been substantial interest in deception in AI ([Masters, Smith, et al. 2021](#); [Ş. Sarkadi, Panisson, et al.](#)

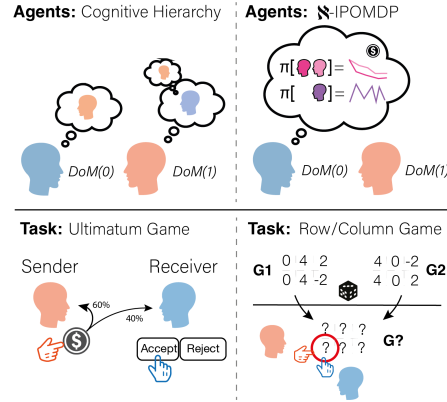


Fig. 1. **Paper overview: (Cognitive Hierarchy:)** We model agents with finite recursive opponent modelling with different Depth of Mentalising (DoM). In the classic model, the player at DoM(0) is at the mercy of the DoM(1) partner given that the DoM(0) player cannot form nested beliefs about their opponent. **(N-IPOMDP:)** The DoM(0) can overcome its recursive limitations by augmenting the classic model. We augment agents' inference processes with an anomaly detection mechanism that allows a self to detect deceptive others by matching expectations with observations. **(IUG:)** Agents with different degrees of DoM interact in the iterated ultimatum game (IUG). In the IUG, on each trial, a Sender offers a split of endowed money, and the Receiver decides whether to accept. If the Receiver decides to reject the offer, both players get 0. **(Row/column game:)** Agents interact in an iterated Bayesian zero-sum game. In this game, nature selects a payoff matrix (G^1 or G^2). Only the row player knows which payoff matrix is sampled and uses this information to their advantage. The column player makes inferences about the payoff matrix from the row player's behaviour.

2019; Ş. Sarkadi, Rutherford, et al. 2021; Savas et al. 2022), for which our work may serve as a blueprint for systems that regulate and possibly prevent AI agents from deceiving other AI or human agents.

1.1 Previous Work

The recursive structure of ToM implies that agents can not make correct inferences about other agents with DoM levels that are higher than their own. Modelling a more sophisticated agent would require a self to model itself from the perspective of another—outside of the DoM hierarchy—and thus, at least in principle, to violate key logical principles (Pacuit and Roy 2017). Such a restriction is not unique to ToM-based models but is evident in general bounded rationality environments (Nachbar and Zame 1996). The relation between deception and recursive ToM has been explored by Oey et al. (2023); in their work, however, agents are aware of the fact the others may deceive them and hence decide whether the others are lying or not. Nevertheless, humans take action under the belief that others are being deceptive, even in the absence of a fully fledged understanding of why. In our work, we consider the case in which agents cannot reason about the possibility of deceptive behaviour, but they can detect a deviation from *expected* behaviour. This is somewhat analogous to off-the-equilibrium-path in Perfect Bayesian Equilibrium (PBE) — certain behaviours (policies) should not appear in the interaction. If unexpected behaviour is observed, then it signals to the agent that something in the assumptions about behaviours is wrong.

The problem of interacting with unmodeled others was introduced in ad-hoc teamwork (Mirsky et al. 2022; Stone et al. 2010). In a cooperative setting (Hu et al. 2021), an agent needs to learn a policy (behaviour) to interact successfully with other agents. ToM allows successful interaction with others (Weerd et al. 2022). It is, however, assumed that the others are properly modelled by the acting agent.

Information-theoretic methods have previously been used for both deception modelling and deception detection. For example, Kopp et al. (2018) introduced a deception modelling framework to study deceptive diffusion along

with cooperation in populations exposed to fake news. Others have looked at detecting masquerading by inferring deception in the context of intrusion detection. For example, [Evans et al. \(2007\)](#) used the MDLCompress algorithm to identify intruders disguised as legitimate users. [Maguire et al. \(2019\)](#) suggests that humans apply a typical set-like mechanism to identify a non-random pattern. However, this is very specific to random behaviour. In this work, we advance this concept to explore several types of deviation away from expected behaviour.

Behaviour-based Intrusion Detection Systems (IDS) methods were proposed by [Pannell and Ashman \(2010\)](#) and [Peng et al. \(2016\)](#). In these systems, the system administrator monitors the behaviour of users to decide and respond to malicious behaviour. However, unlike our proposed method, these systems often require labelled data, making them susceptible to an aware adversary who knows how to avoid detection.

In the context of POMDP, several goal recognition methods have been proposed, for example ([Le Guillarme et al. 2016](#); [Ramírez and Geffner 2011](#)). While these methods assume that the observed agent may be malign, they assume that (a) the observer can invert the actions to make inferences about the malign intent, and (b) they use the likelihood of the observed behaviour to recognise this intent. Here, we show how likelihood-based inference is used *against* the inferring agent, and propose a mechanism that detects deviation from expected behaviour, flagging malevolent agents without making inferences about their goals.

[Yu et al. \(2021\)](#) explores how an agent might adapt to higher DoM opponents by learning the best response from experience. However, their mechanism lacks a model to detect when the opponent's DoM level exceeds the agent's DoM level, which is a necessary step to prompt the agent to retrain its model. On the other hand, ([Piazza et al. 2023](#)) introduces a communicative MARL framework and shows that using ToM to defend against deceptive adversaries has several limitations that are determined by the multi-agent context in which communication is exchanged. For instance, the authors show that deceivers thrive in MARL settings where they can leverage biases such as static play, learning biases, and irrationality. Moreover, their model assumes that the victim can utilise their ToM to make inferences about the intention of the deceiver—distinguishing deceptive behaviour from non-deceptive (up to some limitation described in the paper). In this work we assume that the victim is deceived by an opponent with greater ToM capacity, depriving the victim of the ability to use its ToM to detect deception.

1.2 Theory of Mind in Multi-agent RL

In a partially observed single-agent RL problem (POMDP; ([Kaelbling et al. 1998](#))), an agent is uncertain at time t about the exact state of the environment, $s^t \in S$. It performs an action $a^t \in A$ and receives an observation $o^t \in O$ which depends on both of these: $p(o^t|a^t, s^t)$. Assume that the dynamics of the environment are known to the agent (as we consider model-based RL) and that they are governed by a transition probability: $p(s^t|s^{t-1}, a^{t-1}) \equiv T(s^t, a^{t-1}, s^{t-1})$. The agent makes an inference about the unknown state from the history, $h^{t-1} = \langle a^0, o^0, \dots, a^{t-1}, o^{t-1} \rangle$, in a Bayesian manner: $b(s^t) = p(s^t|h^{t-1})$. This can be expressed recursively as:

$$b(s^t) = p(s^t|h^{t-1}) \propto P(o^{t-1}|a^{t-1}, s^{t-1}) \sum_{s^{t-1} \in S} T(s^t, a^{t-1}, s^{t-1}) b(s^{t-1}) \quad (1)$$

The agent's utility is a function of the environment and its actions: $u : A \times S \rightarrow \mathbb{R}$. An agent's goal is to maximize its long-term discounted utility: $\sum_{t=1}^T \gamma^t u^t \exp[-t \log(\gamma)]$, where $\gamma \in [0, 1]$ is the discount factor. In a POMDP, the optimal action-value function takes a particular action, and reports the sum of the expected immediate utility for that action at a belief state, and the expected long run discounted utility if the best possible action is subsequently taken at each subsequent step. It is defined recursively as:

$$Q^*(a^t, b(s^t)) = E_{s^t \sim b(s^t)} [u^t(a^t, s^t) + \gamma \sum_{o^t \in O} P(o^t|a^t, s^t) \times \sum_{s^{t+1} \in S} T(s^{t+1}, a^t, s^t) \max_{a^{t+1}} \{Q^*(a^{t+1}, b(s^{t+1}))\}] \quad (2)$$

The agent's policy is either a deterministic or stochastic function of these Q-values. One such stochastic policy is the SoftMax policy—a distribution over the actions with a known inverse temperature $\mathcal{T}$:

$$\pi(a^t|b(s^t)) = P(a^t|b(s^t)) \propto \exp \frac{Q^*(a^t, b(s^t))}{\mathcal{T}} \quad (3)$$

In this work we assume that agents follow this policy.¹

In multi-agent RL (MARL), multiple RL agents interact with the same environment. We call the agents μ and ν , and index their actions accordingly.² The interaction implies that the environment changes as a function of other agents' actions as well $p(s^t|a_\mu^{t-1}, a_\nu^{t-1}, s^{t-1}) \equiv T(s^t, a_\mu^{t-1}, a_\nu^{t-1}, s^{t-1})$. Since the agent's utility is a function of the environmental state, it now becomes a function of other agents' actions, too. We highlight this dependency via reformulation of μ 's utility: $u_\mu : A_\mu \times A_\nu \times S \rightarrow \mathbb{R}, u_\mu^t = f(a_\mu^t, a_\nu^t, s^t)$.

Given this coupling, agents in MARL are motivated to predict the behaviour of others (Wen et al. (2019)). Driven by reward maximisation, and since its reward depends on ν 's behaviour, μ may *simulate* ν 's reaction to μ 's action and compute the expected utility of that action:

$$E_{a_\nu^t} [u_\mu^t(a_\mu^t, a_\nu^t)] = \sum_{a_\nu^t} \hat{P}(a_\nu^t|a_\mu^t) u_\mu^t(a_\mu^t, a_\nu^t) \quad (4)$$

Where $\hat{P}(a_\nu^t|a_\mu^t)$ is μ 's simulation of ν . Thus, given a policy, π_μ , agent μ may utilise this simulator to compute the discounted cumulative expected utility of this policy:

$$E(u_\mu|\pi_\mu) = \sum_{t=0}^T E_{a_\nu^t} [u_\mu^t(a_\mu^t, a_\nu^t)] \exp [t \log(\gamma)] \quad (5)$$

This is conceptually the mechanism governing the standard Nash equilibrium (NE). However, NE requires common knowledge (implying an arbitrarily deep level of nested knowledge) of the world and of others' behaviour (Chwe 2013). This assumption is easily revoked in multiple social interactions, where agents typically hold private and public information.

Bayesian games (Zamir 2020) model incomplete information encounters and are used to model social interaction. Formally, in a Bayesian game, agents do not have full information about the environment. This uncertainty may concern the state of the world ("Is there milk in the refrigerator?"), the intentions of other agents ("Friend or Foe?") (Littman et al. 2001) or another's modus operandi ("how will Joe react to me doing this?"). Beliefs about unknown variables govern an agents' decision-making and are updated in a Bayesian manner, in a way similar to a single-agent POMDP. These beliefs may include beliefs about other agents' beliefs—a distribution over distributions (Silva et al. 2024). Furthermore, agents can form beliefs about these recursive beliefs, i.e., a distribution over the distribution of distributions.

This recursive reasoning is known as Theory-of-Mind (ToM). ToM is the ability to ascribe intentionality to others. This trait is considered one of the hallmarks of human cognition (Premack and Woodruff 1978), and has been suggested to be vital to complex human behaviour, from effective communication (Frank and Goodman 2012; Goodman and Frank 2016) to deception (Alon, Schulz, Rosenschein, et al. 2023a; Ş. Sarkadi, Panisson, et al. 2019). Due to the pivotal role ToM plays in the human capacity to interact socially, it has elicited interest from the AI community (Cuzzolin et al. 2020; Q. Wang et al. 2021).

The IPOMDP framework (Gmytrasiewicz and Doshi 2005) combines ToM and POMDP, allowing us to model and solve these recursive beliefs models as a POMDP. While other models formalise recursive ToM (DoM), such as CHASE (Buergi et al. 2024), hypergames (Bennett 1980) or the aforementioned RSA model (Goodman and Frank 2016), the IPOMDP generalises each of these frameworks into a flexible process that can be adapted to

¹Acknowledging the discrepancy with the definition of the optimal Q^* values.

²We discuss two agents, but the same principles apply to a larger number.

several social contexts. Naturally, while the principles presented in this work may be incorporated into any existing DoM model, we opted for the IPOMDP for its generality and wide use (Alon, Schulz, Dayan, et al. 2023; Hula, Vilares, et al. 2018; Rusch et al. 2020). In this model, an agent's *type* (Harsanyi 1968; Westby and Riedl 2022), θ_μ , is a combination of the agent's *persona* (utility function, sensor capacity, etc.), denoted by ψ_μ , and the agent's *beliefs* about the world $b_\mu(\cdot)$: $\theta_\mu = \langle \psi_\mu, b_\mu(\cdot) \rangle$. It is assumed that agents are fully Bayesian, meaning that their uncertainty is epistemic and can be reduced to correctly identify the *type* of agents with a lower DoM. Misplaced beliefs may only result from deceptive behaviour as explained in the following sections, and not from misplaced priors or insufficient support.

Inference with ToM. Beliefs can be limited to environmental uncertainty alone (reducing the problem to a POMDP) or they may also include beliefs from an intentional agent. The depth of belief recursion, $k \in [-1, \infty)$, is known as the agent's Depth of Mentalising (DoM). The Interactive State (IS) augments the concept of state in POMDP to account for the multi-variable nature of the problem (both environmental uncertainty and opponent uncertainty) $is_{\mu_k}^t = \langle s^t \times \theta_{\nu_{k-1}}^t \rangle$, where $\theta_{\nu_{k-1}}^t = \langle \psi_{\nu_{k-1}}, b_{\nu_{k-1}}(is_{\nu_{k-1}}^t) \rangle$ is the *type* of the DoM($k-1$) agent—its persona and nested beliefs. This representation illustrates the recursive structure of ToM as it is possible to replace $is_{\nu_{k-1}}^t$ with $\langle s^t \times \langle \psi_{\mu_{k-2}}, b_{\mu_{k-2}}(is_{\mu_{k-2}}^t) \rangle \rangle$ —revealing the hierarchical belief structure of ToM, in which μ_k reasons about ν_{k-1} reasoning about μ_{k-2} and so on.

During the interaction, the acting agent μ observes the actions (either sequentially or simultaneously) of ν (that is μ 's observations are ν 's actions). In this work, assume that there is no environmental uncertainty, focusing on the strategic behaviour arising from uncertainty about the opponent (we refer the reader to (Gmytrasiewicz and Doshi 2005) for a full description of belief update in IPOMDP). The resulting Bayesian updated beliefs about ν 's type are:

$$b_{\mu_k}(is_{\mu_k}^t) = p(\theta_{\nu_{k-1}}^t | h^{t-1}) \propto \sum_{\theta_{\nu_{k-1}}^{t-1} \in \Theta_{\nu_{k-1}}} P(a_{\nu_{k-1}}^{t-1} | \theta_{\nu_{k-1}}^{t-1}, a_{\mu}^{t-1}) p(\theta_{\nu_{k-1}}^{t-1} | h^{t-2}) \quad (6)$$

Since μ 's actions affect ν 's beliefs, but ν 's persona (ψ_ν) is assumed immutable, we expand Eq. 6 to distinguish between inference about temporal changes to ν 's belief (potentially about μ 's type) and inference about ν 's unalterable persona:

$$p(\theta_{\nu_{k-1}}^t | h^{t-1}) = p(\langle b_{\nu_{k-1}}^t, \psi_\nu \rangle | h^{t-1}) \propto \sum_{\psi_\nu \in \Psi_\nu} (P(a_{\nu_{k-1}}^{t-1} | \langle b_{\nu_{k-1}}^{t-1}, \psi_\nu \rangle, a_{\mu}^{t-1}) \times P(b_{\nu_{k-1}}^t | b_{\nu_{k-1}}^{t-1}, a_{\mu}^{t-1})) p(\theta_{\nu_{k-1}}^{t-1} | h^{t-2}) \quad (7)$$

The Cognitive Hierarchy. At the core of recursive models are subintentional agents (sometimes referred to as DoM(-1)). These agents follow reactive, typically model-free and myopic policies, treating other agents' actions as environmental features rather than strategic choices. Their optimal policy depends only on their own policy and the interaction history:

$$\pi_{\mu-1}^t(\psi_\mu, a_{\nu}^{t-1}) = P(a_{\mu}^t | \psi_\mu, a_{\nu}^{t-1}) \quad (8)$$

The DoM(0) models other agents as having DoM(-1), implying that these agents model the opponent and environment separately. Plugging Eq. 8 into Eq. 6 yields the DoM(0) belief update, which is a simple Bayesian IRL (Ramachandran and Amir 2007):

$$b_{\nu_0}(\theta_{\nu_0}^t) = p(\psi_\mu | h^{t-1}) \propto \sum_{\psi_\mu \in \Theta_{\mu-1}} P(a_{\mu}^{t-1} | \psi_\mu, a_{\nu}^{t-1}) p(\psi_\mu | h^{t-2}) \quad (9)$$

Using these beliefs the DoM(0) computes the state-action values (Q-values):

$$Q_{\nu_0}^*(a_{\nu}^t, \theta_{\nu_0}^t) = E_{a_{\mu}^t \sim \pi_{\mu-1}(\psi_\mu, a_{\nu}^{t-1})} [u_{\nu}^t(a_{\nu}^t, a_{\mu}^t) + \max_{a_{\nu}^{t+1}} [Q_{\nu_0}^*(a_{\nu}^{t+1}, \theta_{\nu_0}^{t+1})]] \quad (10)$$

where $u_v^t(a_v^t, a_\mu^t)$ is v 's utility at time t . The DoM(1) models others as DoM(0) agents, making inferences about their beliefs as well:

$$b_{\mu_1}(\theta_{\mu_1}^t) = p(\psi_v \times b_{v_0}(\theta_{v_0}^t) | h^{t-1}) \propto P(a_v^{t-1} | \psi_v, b_{v_0}(\theta_{v_0}^{t-1}), a_\mu^{t-1}) \times P(b_{v_0}(\theta_{v_0}^t) | b_{v_0}(\theta_{v_0}^{t-1}), a_\mu^{t-1}) p(\psi_v, b_{v_0}(\theta_{v_0}^{t-1}) | h^{t-2}) \quad (11)$$

This trait allows it to reason about changes in the beliefs of DoM(0) and use this ability to its benefit, as presented next. Similarly to the DoM(0), the DoM(1) computes the Q -values of its actions and acts based on these values.

1.3 Planning with ToM

Theory of Mind allows the acting agent to simulate other agents, taking their perspective (Eq. 4). Formally, this is related to the concept of sequential rationality in extensive-form Bayesian games (Fudenberg and Tirole 1991; Harsanyi 1968). In the current work, the acting agent maintains a belief system over the opponent's hidden 'types' — represented by their DoM level and persona — and computes optimal policies modulo this mental model. This capacity empowers agents with high DoM to shape the behaviour ((Jaques et al. 2019; Kim et al. 2022)) of lower DoM agents to their benefit. We formally describe this process before illustrating how the utility function of the acting agent is integrated into its policy computation.

Using its nested mental model of v , denoted $\hat{\theta}_{v_{k-1}}$, the acting agent μ_k computes v 's optimal policy, given its model of v 's belief and persona:

$$\hat{\pi}_{v_{k-1}}^t(\hat{\theta}_{v_{k-1}}^t) = p(a_v^t | \hat{b}_{v_{k-1}}(\hat{\theta}_{v_{k-1}}^t), \psi_v) \quad (12)$$

where $\hat{\pi}_{v_{k-1}}^t(\cdot)$ denotes the simulated policy of a nested opponent model and $\hat{b}_{v_{k-1}}(\hat{\theta}_{v_{k-1}}^t)$ denoted the nested belief about a nested self model—the belief μ_k ascribes to v_{k-1} about μ , as modelled by v . From Eq. 12 μ computes a distribution of future actions conditioned on a given policy:

$$p(a_v^{t+1:T} | \hat{\theta}_{v_{k-1}}^t, \pi_{\mu_k}^t) = \prod_{i=1}^{T-t} \sum_{a_\mu^{t+i-1}} p(a_v^{t+i} | \hat{\pi}_{v_{k-1}}^{t+i}(\hat{\theta}_{v_{k-1}}^{t+i})) p(\hat{b}_{v_{k-1}}(\hat{\theta}_{v_{k-1}}^{t+i-1}) | a_\mu^{t+i-1}) P(a_\mu^{t+i-1} | \pi_{\mu_k}^{t+i-1}) \quad (13)$$

where $\pi_{\mu_k}^t$ is μ 's revised policy after observing a_v^{t-1} and updating its beliefs about $\hat{\theta}_{v_{k-1}}^t$.

The interplay between μ 's utility and v 's behaviour affects μ 's value function. The *value* function of a policy π_{μ_k} (Sutton and Barto 2018) is defined as the expected cumulative utility of μ_k if it samples its actions from it ($a_\mu^t \sim \pi_{\mu_k}^t(\psi_\mu, b_{\mu_k}(\theta_{v_{k-1}}^t))$):

$$V^{\pi_{\mu_k}^t}(b_{\mu_k}(\theta_{v_{k-1}}^t)) = E_{\pi_{\mu_k}^t} [E_{a_v^t \sim \hat{\pi}_{v_{k-1}}^t(\psi_v, b_{v_{k-1}}(\theta_{v_{k-1}}^t))} [u_\mu(a_\mu^t, a_v^t) + \gamma V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))]] \quad (14)$$

Changes in the policy are propagated to the value function through two components - an *immediate* effect and a *gradual* effect caused by changes in v 's behaviour. We illustrate these effects to highlight the role of ToM in optimal (at least from the DoM(k) agent's perspective) policy (the full computation is presented in App. 6.1):

$$\frac{\partial V^{\pi_{\mu_k}^t}(b_{\mu_k}(\theta_{v_{k-1}}^t))}{\partial \pi_{\mu_k}^t} = \frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^t} [u_\mu^t(a_\mu^t, a_v^t)]]}{\partial \pi_{\mu_k}^t} + \gamma \frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^t} [V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))]]}{\partial \pi_{\mu_k}^t} \quad (15)$$

The immediate effect (first RHS argument of Eq. 15) measures the change of μ 's current utility as a function of its action (assuming simultaneous actions, but this can be adapted to sequential actions, too):

$$\frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^t} [u_\mu^t(a_\mu^t, a_v^t)]]}{\partial \pi_{\mu_k}^t} = \sum_{a_\mu^t} \frac{\partial P(a_\mu^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} \bar{u}(a_\mu^t) \quad (16)$$

Where $\bar{u}(a_\mu^t) = E_{\pi_{v_{k-1}}^t} [u_\mu^t(a_\mu^t, a_v^t)]$ is the expected utility for μ from playing action a_μ^t averaged over v 's policy.

The gradual effect $(t + 1)$ describes the long-term effect on μ 's value, stemming from v 's adapted behaviour to μ 's action ((Na et al. 2021; Siu 2022)):

$$\frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^{t+1}} [V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))]]}{\partial \pi_{\mu_k}^t} = \sum_{a_v^t} P(a_v^t | \pi_{v_{k-1}}^t) \left[\sum_{a_\mu^t} \frac{\partial P(a_\mu^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} \times V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1})) + P(a_\mu^t | \pi_{\mu_k}^t) \times \frac{\partial V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t} \right] \quad (17)$$

The term $\frac{\partial P(a_\mu^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))$ relates to *policy gradient* (Sutton, McAllester, et al. 1999). The second term corresponds to *opponent shaping* (Foerster et al. 2018). Unlike the work of Foerster et al. (2018), the ToM opponent shaping takes an *indirect* path, whereby the shaping agent induces behavioural change by affecting others' beliefs. This agency is expressed through the following equation (see App. 6.1 for full derivation):

$$\frac{\partial V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t} = \frac{\partial V^{\pi_{\mu_k}^{t+1}} (b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{v_{k-1}}^{t+1}} \frac{\partial \pi_{v_{k-1}}^{t+1}}{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})} \frac{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})}{\partial \pi_{\mu_k}^t} \quad (18)$$

First, μ 's action changes v 's belief: $\frac{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})}{\partial \pi_{\mu_k}^t}$, computed via derivation of Equation 6. For example, if v 's beliefs are parametric, this change is computed via the *influence* function (Koh and Liang 2017).

Next, the changes in v 's belief affect its Q-value computation and hence its policy $\frac{\partial \pi_{v_{k-1}}^{t+1}}{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})}$. Moreover, the long term effect of shaping v 's beliefs in favour of μ is propagated into future steps through v 's updated beliefs (as illustrated in Sec. 2.1.1).

In a cooperative game, shaping is often the revelation of unknown information or a nudge, aimed to improve others utility (Jaques et al. 2019), as it is assumed that the agents' goal is shared and therefore there is no incentive for μ to hide or falsify information (Devaine et al. 2014). However, this is not the case in non-cooperative games, where often agents can gain from disclosing information or providing signals that cause the observer to form *false beliefs* (Alon, Schulz, Dayan, et al. 2023; Alon, Schulz, Rosenschein, et al. 2023a); this concept is presented in the next section.

2 Deception with ToM

Deception is defined as the deliberate causation of a false belief in the mind of a target with an ulterior motive (Kopp et al. 2018; Ş. Sarkadi, Panisson, et al. 2019; Ward 2023). However, assessing the success of deception is non-trivial; there is often no guarantee that a victim's action was induced by the intended false belief or if the victim acted independently (Masters and Sardina 2017). We formalize and rigorously operationalize this in our framework, we propose an axiomatic construction of deception based on sequential rationality and Depth of Mind (DoM). We define a deceptive policy $\pi_\mu^\dagger$ through four necessary conditions: Incentive Compatibility, Epistemic Manipulation, Behavioural Causality, and Cognitive Dominance.

Axiom 1: Incentive Compatibility (The Ulterior Goal). A rational agent will only engage in deception if it yields a higher expected utility than honest behaviour. Formally, let $\pi_\mu^\dagger$ denote a policy that knowingly aims to install a false belief. The agent μ will only choose $\pi_\mu^\dagger$ if it is at least as valuable as any other policy:

$$V^{\pi_\mu^\dagger} (b_{\mu_k}(\theta_{v_{k-1}})) \geq V^{\pi_\mu} (b_{\mu_k}(\theta_{v_{k-1}})) \quad (19)$$

This inequality satisfies the motivation established in Eq. 15: μ is incentivised to “affect” v specifically because the resulting behaviour generates excess utility.

Axiom 2: Epistemic Manipulation (False Beliefs). Deception requires the transmission of a signal intended to induce a divergence between the victim’s belief and the deceiver’s true parameters. By false belief, we mean a belief that μ believes (knows) is false (Ward 2023), effectively targeting the victim’s *perception* (Eger and Martens 2017). A policy $\pi_\mu^\dagger$ for a deceiver of type $\theta_\mu^\dagger = \langle \psi_{\mu^\dagger}, b_{\mu^\dagger}(\cdot) \rangle$ installs *false beliefs* if, during a deceptive phase $T_D = \{t_1, \dots, t_D\} \subseteq \{0, \dots, T\}$, the victim assigns higher probability to an incorrect type $\theta_\mu^\star$ than the true type $\theta_\mu^\dagger$:

$$\exists T_D : \forall t \in T_D : P_v(\theta_\mu^\star | h^t) \geq P_v(\theta_\mu^\dagger | h^t) \quad (20)$$

Given the cognitive hierarchy prohibition on modelling higher DoM-level agents’ beliefs, we decompose Eq. 20 to highlight two aspects of the deceptive policy: installing false beliefs about the *persona* ψ , and manipulation of nested beliefs. At any DoM level, the deceiver manipulates the victim to falsely believe that its persona ψ_μ is $\psi_{\mu^\star}$ rather than $\psi_{\mu^\dagger}$:

$$\exists T_D : \forall t \in T_D : P_v(\psi_{\mu^\star} | h^t) \geq P_v(\psi_{\mu^\dagger} | h^t) \quad (21)$$

For example, a deceptive foe (higher DoM) may portray itself as a friend (Adhikari and Gmytrasiewicz 2021), satisfying this axiom by manipulating the probability mass over the persona set.

Recurring through higher levels ($k \geq 2$), we use Eq. 21 to define *false nested beliefs*. Crucially, by definition, the victim (DoM(k)) cannot fully represent the deceiver’s beliefs, as they include $(k + 1)$ beliefs, implying that the victim’s beliefs about the deceiver’s *type* are wrong. Nonetheless, since both agents share at least $(k - 2)$ beliefs (for example, if $k = 3$, they share both level 1 and level 0 beliefs) — the manipulation of these shared beliefs is part of the deception. That is, as part of the ploy, the higher DoM deceiver can manipulate the victim’s beliefs about μ ’s (the deceiver) beliefs about v (the victim). For example, the deceiver may aim to not only cause the victim falsely believe that they are friendly, but also to induce false (nested) beliefs that they themselves (μ) believe that v is a friend too. Formally, let $p_\mu^\dagger(\psi_v | h^t)$ denote μ ’s *true* belief about v ’s persona. Similarly, we define the *false* (or wrong) belief as $p_\mu^\star(\psi_v | h^t)$. Using these notations, and the $\hat{\cdot}$ notation for estimated nested beliefs, we define v ’s *false nested belief* about μ ’s beliefs as:

$$\exists T_D : \forall t \in T_D : P_v(\hat{p}_\mu^\star(\psi_v | h^t)) \geq P_v(\hat{p}_\mu^\dagger(\psi_v | h^t)) \quad (22)$$

Axiom 3: Behavioural Causality (Regret). It is not sufficient for v to merely hold a false belief; that belief must “tip the scales” (S. Sarkadi 2021) to induce a behaviour that is suboptimal for the victim but beneficial for the deceiver. We measure this via *regret* (Blum and Mansour 2007; Jin et al. 2018), defined here as the difference between the victim’s chosen action (based on false belief $b_v^{t,\dagger}$) and the optimal action they would have taken had they possessed the cognitive capacity (DoM($k + 1$)) to see through the bluff:

$$Reg_{v_{k-1}}^t = E_{a_v^t \sim \pi_{v_{k-1}}^t(b_v^{t,\dagger})} u_v^t(a_\mu^t, a_v^t) - E_{a_v^t \sim \pi_{v_{k+1}}^t(b_v^t)} u_v^t(a_\mu^t, a_v^t) \quad (23)$$

Since the victim cannot compute this counterfactual, we approximate the discrepancy by comparing the realized reward against the victim’s expected reward under their (manipulated) belief. Using Eq. 12, the victim’s expected reward for a given belief state is:

$$E(\hat{r}_v^t) = \sum_{\hat{\theta}_{\mu_{k-2}}} \hat{r}_v^t(\hat{\theta}_{\mu_{k-2}}) P_{v_{k-1}}^{t-1}(\hat{\theta}_{\mu_{k-2}}) \quad (24)$$

The measurable estimator for the success of the deception is thus:

$$Reg^t = r_v^t - E(\hat{r}_v^t) \quad (25)$$

In zero-sum games, a negative Reg^t for the victim corresponds to the exact advantage gained by the deceiver.

Axiom 4: Cognitive Dominance (Avoiding Detection). Finally, for deception to persist, the victim must remain unaware of the manipulation. This necessitates a cognitive asymmetry where $DoM(\mu) > DoM(\nu)$. Because ν is limited to making inferences about lower-level DoM agents, it cannot straightforwardly interpret the deceiver's actions as deceptive without violating logical principles of self-reflection (Pacuit and Roy 2017). We conclude that ToM is the necessary mechanism to satisfy these axioms. It enables the simulation required for Axiom 2, the policy optimization for Axiom 1, and the exploitation of the cognitive gap defined in Axiom 4 to induce the regret defined in Axiom 3.

2.1 Illustration in Bayesian Games

We illustrate deceptive behaviour in a mixed-motive and a zero-sum game. Since the main results and findings are similar, we present the mixed-motive game results in the main body and the zero-sum game results in Appendix 6.4. In both cases, we illustrate how higher DoM agents utilise nested models to simulate lower DoM agents, and how this nesting is used to maximise reward. We discuss the details of the deception concerning Eq. 15 and show how the policy balances short- and long-term rewards. We conclude that ToM-based deception follows a pattern of installing false beliefs, followed by the execution of undetected strategic behaviour.

2.1.1 Mixed-Motive Game. The iterated ultimatum game (IUG) (Alon, Schulz, Dayan, et al. 2023; Camerer 2011) (illustrated in Fig 1(Ultimatum Game)); is a repeated Bayesian mixed-motive game (Alon, Schulz, Dayan, et al. 2023). Briefly, the IUG is played between two agents—a sender and a receiver. We will consider how the sender might try to deceive the receiver, and so designate them as μ and ν respectively, following the convention established above. On each trial t of the game, the sender gets an endowment of 1 monetary unit and offers a (for convenience, discretized) percentage of this to the receiver: $a_\mu^t \in \{0, 0.1, 0.2 \dots, 1\}$, $t \in [1, T]$ (we used $T=12$). If the receiver *accepts* the offer ($a_\nu^t = 1$), the receiver gets a reward of $r_\nu^t = a_\mu^t$ and the sender a reward of $r_\mu^t = 1 - a_\mu^t$. Alternatively, the receiver can *reject* the offer ($a_\nu^t = 0$), in which case both players receive nothing. In this game, agents need to compromise (or at least consider the desires of each other) in order to maximise the reward that they can achieve—this makes the IUG a useful testbed for assessing strategic mentalising.

We simulated senders with two DoM levels: $k \in [-1, 1]$ interacting with a DoM(0) receiver. The DoM(-1) senders came in three behavioural types: one emits uniformly random offers (and so is not endowed with Q-values); the other two have utility functions that are characterized by a threshold, $\psi_\mu \in \{0.1, 0.5\}$: $u_\mu^t(a_\mu^t, a_\nu^t, \theta_\mu) = (1 - a_\mu^t - \psi_\mu) \cdot a_\nu^t$. We call these senders the threshold senders. The threshold senders compute Q-values based on their models of the world and select their actions via the SoftMax policy (Eq. 3). The DoM(-1) threshold senders compute the Q-values in the following way—they maintain lower and upper bounds on the viable offer set, L^t, U^t . These bounds are updated:

$$L^t = L^{t-1} \cdot a_\nu^{t-1} + a_\mu^{t-1} \cdot (1 - a_\nu^{t-1}) \quad (26)$$

$$U^t = U^{t-1} \cdot (1 - a_\nu^{t-1}) + a_\mu^{t-1} \cdot (a_\nu^{t-1}) \quad (27)$$

with $L^0 = 0$ and $U^0 = 1$. In this work, we assume that the DoM(-1) agents are extremely myopic, reducing the Q-values to the immediate reward (Alon, Schulz, Dayan, et al. 2023). In turn, these senders' Q-values (in fact, degenerate Q-values) are simply the immediate utility from every action in the range $a_\mu^t \in (L^t, U^t]$:

$$Q_{\mu-1}^t(a_\mu^t; \psi_\mu) = u_\mu^t(a_\mu^t, \psi_\mu) \quad (28)$$

and not computing the values for actions outside of this interval (narrowing the possible actions after each iteration). This assumption can be revoked in future work. The sender will select an offer (via SoftMax policy) from the set of potential offer based on the offer's Q-value. We model these agents as rational agents, meaning they will not engage in a losing interaction. Lacking the option to quit, if the lower bound L^t reaches the sender's threshold ψ_μ we assume that the sender will not update the bounds and will offer a partition that yields it a zero

reward. Since these agents are assumed to lack an opponent model and are often modelled (Gmytrasiewicz and Doshi 2005) as having a uniform distribution over future events, we expect the results to hold.

The DoM(0) receiver infers the behavioural type of the DoM(−1) sender from their action as in Eq. 9. As described above, we assume that the DoM(0) receiver has full knowledge of the potential types of senders ($\Theta_{\mu-1}$) and its uncertainty is limited to which realization is sampled $\psi_{\mu-1} \in \Theta_{\mu-1}$. We leave uncertainty about Θ_{μ} to future work. Given the updated beliefs, they compute the Q-values (Eq. 10), via Expectimax search (Russell 2010) and select an action via a SoftMax policy (Eq. 3) with temperature ($\mathcal{T} = 0.1$; which is common knowledge). Lastly, the DoM(1) sender simulates the DoM(0) receiver’s beliefs and resulting actions during the computation of its Q-values using the IPOMCP algorithm (Hula, Montague, et al. 2015), with full planning horizon.

As hypothesised, agents with high DoM take advantage of those with lower DoM. The complexity of this manipulation rises with the agents’ level. The DoM(0) receiver infers the type of the DoM(−1) sender and uses its actions to manipulate the behaviour of the sender (if possible). Specifically, if the receiver believes they are partnered with an intentional (non-random) sender, they will try to influence the sender to offer more via strategic rejection (as illustrated in Fig. 2, illustrating one simulation). This will continue until either the sender’s threshold is met or until the offers are sufficient given the opportunity cost implied by the planning horizon. Note that this type of behaviour is deceptive according to our definition, as the receiver uses its model of the sender to cause the sender to improve its offers, which improves the receiver’s utility, by planting “false-beliefs” in the DoM(−1) victim’s “mind”. On the other hand, if a receiver believes they are partnered with a random sender, then, since the random sender affords no agency, the optimal policy for the receiver is to accept any offer.

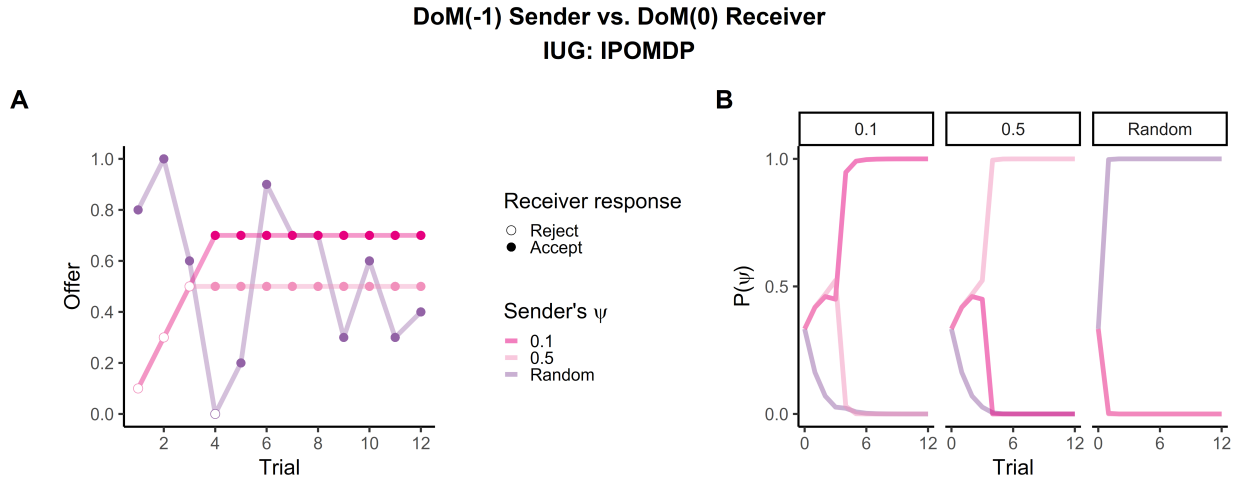


Fig. 2. **DoM(0) vs DoM(−1) in IUG IPOMDP:** (A) The points show offers from the sender to the receiver over all 12 trials, coloured by sender behavioural type (random or utility). Points are shaded white if the receiver rejects the offer. A DoM(0) receiver quickly infers from the initial offers the type of the DoM(−1) sender. The DoM(−1) utility sender’s first offer tells it apart from the random sender as its initial offer is always close to 0 (up to random noise). The DoM(0) policy is a function of its updated beliefs (B). Updated belief probabilities of the receiver when playing with different random or utility senders. DoM(0) receivers are well tuned to detect which type of sender they are partnered with. When engaging with a threshold DoM(−1) sender, the receiver rejects the offers until the sender is unwilling to “improve” its offers, which also corresponds to the certainty of its beliefs

The DoM(1) sender's Q-values are computed in a similar manner to Eq. 10:

$$Q_{\mu_1}^*(a_{\mu}^t, \theta_{\mu_1}^t) = E_{a_v^t \sim \pi_{v_0}(\psi_v, b_{v_0}(\psi_{\mu-1}))} [u_{\mu}^t(a_{\mu}^t, a_v^t) + \gamma \max_{a_{\mu}^{t+1}} [Q_{\mu_1}^*(a_{\mu}^{t+1}, \theta_{\mu_1}^{t+1})]] \quad (29)$$

where, as explained in Section 2, the actions of the DoM(1) sender affect not only its immediate reward, but also the long term reward via the belief manipulation process. The DoM(1) sender sets out to deceive the DoM(0) receiver based on the former's behavioural pattern. Using its capacity to simulate this behaviour fully, the DoM(1) sender acts deceptively by masquerading as a random sender, preying on the lack of agency that the DoM(0) has over the random sender. This policy is depicted in Fig. 3.

The sender begins by making a relatively high offer (Fig. 3(A)). This offer is very unlikely for the threshold DoM(-1) senders, hence the belief update of the DoM(0) receiver strongly favours the random sender hypothesis (Fig. 3(B)) (Axiom 2). Once the receiver's beliefs are misplaced, the sender takes advantage of the statistical nature of the random sender policy—every offer has the same likelihood: $\frac{1}{|\mathcal{A}|}$. This allows the sender to reduce their subsequent offers substantially (Axiom 1), while avoiding “detection” (Axiom 4). Thus, the DoM(1) policy is deceptive policy according to our definition. We note that any deceptive schema depends on the types the deceiver can masquerade to (Eq. 20). In particular, the existence of the random sender enables the DoM(1) sender to act in a way that benefits it and installs false beliefs in the receiver's mind. Even with the lack of a noisy agent, as long as there are other types for which the behaviour of the victim is more beneficial than the policy against the deceiver's true type—deception will occur.

We measure the various metrics of deception during the interaction between the DoM(1) sender and DoM(0) receiver: false beliefs (Eq. 20) and deviation from expected reward (Eq. 25). The values of these are shown in Fig. 3(B,C). First, the actions of the DoM(1) sender causes the DoM(0) receiver to form false beliefs, as illustrated in Fig. 3(B). In this case the receiver's beliefs are misplaced, as: $P(\psi_{\mu} = \text{random}) > P(\psi_{\mu} \in \{0.1, 0.5\})$, throughout the interaction ($t_D = T$). The expected reward from the DoM(0) receiver is computed via Eq. 24 and presented in Fig. 3(C) as the striped bars. These can then be compared to the observed reward (lower, non striped bars). This figure shows that, on average, the DoM(1) sender retains between 40% ($\psi = 0.1$) to 70% ($\psi = 0.5$) more reward than the receiver expects without being detected or causing the DoM(0) to change its behaviour.

The deceptive behaviour of the DoM(1) sender in the IUG task illustrates the main steps of deception with ToM, a pattern that is not unique to the IUG game, but has also been reported in other work, for example, Alon, Schulz, Rosenschein, et al. (2023a). The deceiver is playing a gambit, incurring some cost to itself, at the expense of installing false beliefs which it later exploits. The victim's inability to reason about this complex behaviour causes it to act according to the the deceiver's plan.

3 Overcoming Deception with Limited Computational Resources

The central idea of this paper is that despite limitations in opponent modelling, the victim can resist and detect deception. This is realised by assessing the (mis-)match between the *expected* (based on the victim's lower DoM) and *observed* behaviour of the opponent—a form of heuristic behaviour verification, or prediction error. For example, consider a parasite, masquerading as an ant to infiltrate an ant colony and steal food. While its appearance may mislead the guardian ants, its behaviour—feasting instead of working—should trigger an alarm. Thus, even if the victim lacks the cognitive power to characterise or understand the manipulation, they can detect the strategy through a form of conduct violation.

If the victim detects such a discrepancy, they can conclude that the observed behaviour is generated by an opponent that lies outside their world model. In principle, there is a wealth of possibilities for mismatches, including incorrect prior distributions over parameters governing the DoM(-1) behaviour or the utilities. Here we assume that the only source of poor model capacity is the limited DoM level. For a level k agent, we denote the set of modelled levels as $\Theta_k \subset \Theta$, and the set of unmodelled levels as $\Theta_{\mathcal{N}}$. For a DoM(k) agent, all the possible

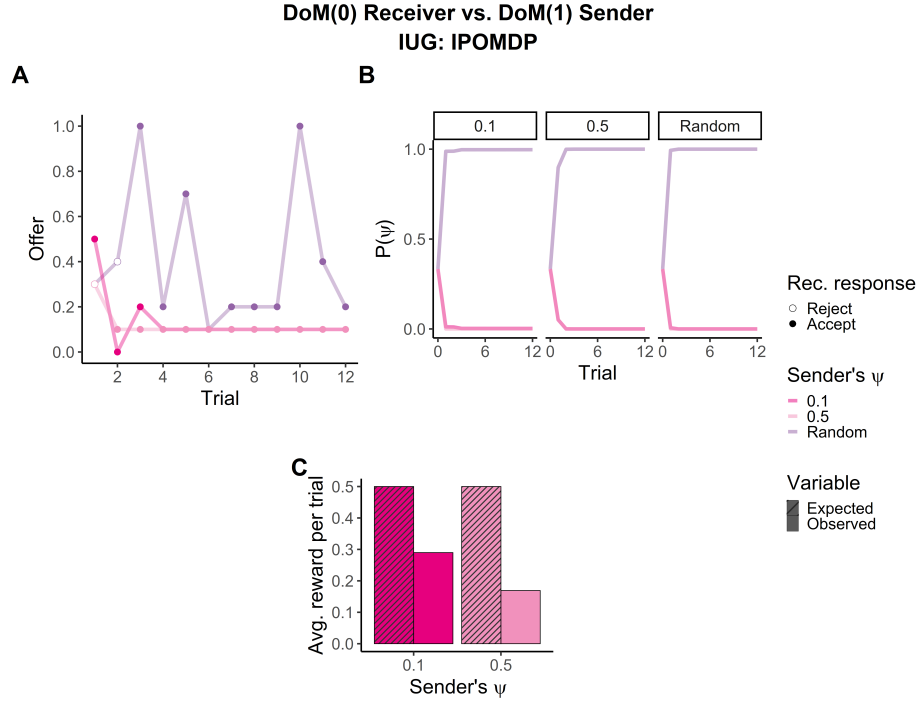


Fig. 3. **Illustration of deception in IUG:** (A) Points show offers from the sender to the receiver over all 12 trials, coloured by sender behavioural type (random or utility). Points are shaded white if the receiver rejects the offer. The DoM(1) acts in a deceptive way to masquerade itself as a random sender, hacking the DoM(0) Bayesian IRL. It starts with a relatively high first offer, and then decreases sharply. (B) Updated belief probabilities of the receiver when playing with different random or utility senders. DoM(0) receivers are poorly tuned to detect the type of DoM(1) sender with which they are partnered, mistaking all actions as if they came from the random sender. This stratagem employed by the sender exploits the pitfall of Bayesian inference used by the DoM(0)—the likelihood that any offer sequence is equal for the random sender. (C) Comparing the expected reward from the DoM(0) receiver's perspective $E(\hat{r}_v^t)$ (striped bars) to the observed reward (non striped bars) each trial. This measures the advantage of the deceiver's policy, reflecting the deceiver's ability to increase its reward at the expense of the victim.

DoM($k - 1$) models are in Θ_k . Thus, a mismatch between the observed and expected behaviour implies that the unknown agent has DoM level *different* from ($k - 1$). This is a pivotal concept, allowing the agent to engage with a known, yet unmodeled, opponent.

Once the behavioural mismatch has been identified, the victim has to decide how to act. Defensive counter-deceptive behaviour should consider the observation that an out-of-model deceiver can predict the victim. Thus, the victim's policy could be aimed at hurting the deceiver (at a cost to the victim), to deter them from doing so.

3.1 Detecting and Responding to Deception with a Limited Opponent Model

Detecting abnormal, and potentially risky, behaviour from observed data is related to Intrusion Detection Systems (IDS). This domain assumes that “behaviour is not something that can be easily stole[n]” (Salem et al. 2008). Thus, any atypical behaviour is flagged as a potential intruder, alarming the system that the observed user poses a risk.

Several methods have been suggested to combat a masquerading hacker (Salem et al. 2008). Inspired by these methods, we augment the victim’s inference with an **S**-mechanism. This mechanism, $f(\Theta_k, h^t)$, evaluates the opponent’s behaviour against the expected behaviour, based on the agent’s DoM level and the history. The expected behaviour includes the presumed opponent’s response to the agent’s actions, similar to the simulated policy in Equation 12. The **S**-mechanism returns a binary vector of size $|\Theta_k|$ as an output. Each entry represents the **S**-mechanism’s evaluation per opponent type: $\theta_v \in \Theta_k$. The evaluation either *affirms* or *denies* that the observed behaviour (discussed next) sufficiently matches the expected behaviour of each agent type.

A critical issue is that, as in Liebald et al. (2007), we allow the deceiver to be aware of the detailed workings of the **S**-mechanism, and so be able to avoid detection by exploiting its regularities. This renders many methods impotent for detecting deception. Nonetheless, to gain an advantage from its deception, the deceiver will have to deviate from typical behaviour at some point (otherwise it will just emit the same actions against which the victim is best responding). For example, in the zero-sum game, the deceiver’s behaviour shifts from typical behaviour for a given opponent policy to non-typical behaviour for the same behavioural type. This should alert the victim that they might be engaging with an unmodelled agent. Of more consequence, at least in a non-fully cooperative game, is that the victim will get less cumulative reward than it expects based on Equation 24. Thus, if the victim’s *actual* reward deviates (statistically) from the expected reward, they can estimate that the opponent is not in their world model.

3.1.1 S-Mechanism. The behaviour verification mechanism assesses whether the observed opponent belongs to the set of potential opponents using two main concepts: typical behaviour monitoring and the expected reward. The concept of typical behaviour arises from Information Theory (IT), where the *typical set* is defined as the set of realisations from a generative model with empirical and theoretical frequencies that are sufficiently close according to an appropriate measure of proximity. This measure is often used to compute asymptotic values such as channel capacity. In this work we use it to decide if a finite sequence of observations is a typical behaviour given an opponent model. If an observed trajectory of behaviour does not belong to an expected typical set there is a high probability that the behaviour was not generated by the expected model. The second component confirms that the observed reward is consistent with the expected reward for each partner type.

Strong typicality is often used to compute the asymptotic properties of a communication channel. Each type acts in a certain way, depending on the duration of the interaction and the responses of the primary agent. Thus, the inferring agent can compute a distribution over the responses of each type in its opponent set (Eq. 12). We present two different approaches to compute this size, depending on the nature of the modelled opponent.

We begin with presenting a δ -strong typicality model. Formally, for each time $t \in [0, T]$, let $\hat{F}_{h^t}^t(a'_v)$ denote the empirical likelihood of an opponent action in the corresponding history set h^t , defined as $\hat{F}_{h^t}^t(a'_v) = N(a'_v)/|h^t|$, where $N(a'_v)$ is the number of times the action a'_v appears in the history set. Let $F_{\theta_v}^t(a'_v)$ be the theoretical probability that an agent with type θ_v will act a'_v at time t given history h^{t-1} .

A δ -strongly typical set of a trajectory, for agent with type θ_v , is the set:

$$Y_\delta^t(\theta_v) = \{h^t : |\hat{F}_{h^t}^t(a'_v) - F_{\theta_v}^t(a'_v)| \leq \delta \cdot F_{\theta_v}^t(a'_v)\} \quad (30)$$

The parameter δ governs the size of the set, which in turn affects the sensitivity of the mechanism. It can be tuned using the nested opponent models to reduce false positives. We analyse the sensitivity of the mechanism to various values of δ in the next section. However, since the deceiving agent model is absent from this “training” set, this parameter cannot be tuned to balance true negatives. An additional issue with setting this parameter

is its lack of sensitivity to history length; the distance between the expected behaviour according to history (theoretical) and the empirical (the actual observed behaviour) reduces with t , but is quite high when t is small. We address this issue by making it trial-dependent, denoted by $\delta(t)$. We discuss task-specific details in Appendix 6.4.

However, this condition is only useful when the observations are independent and identically distributed (iid) variables. This holds for a random sender in the IUG task, or the agents depicted in the Row/Column game (Appendix 6.4). In many multi-agent interactions, the actions at time t depend on the history up to that point. For example, as illustrated in Fig. 2, the DoM(-1) threshold sender behaviour (policy) changes each trial. To overcome this issue we present a second computational method to estimate a form of sequential typicality.

Inspired by the work of (N. Wang et al. 2012), we propose a gzip-based (Deutsch 1996) algorithm to compute the sequential typicality set. This is a sampling-based algorithm, depicted in Alg. 2. The core idea is to generate independently sampled trajectories for each simulated opponent policy (Eq. 12), adding $\hat{o}_{v,n}^t \sim \hat{\pi}_{v_{k-1}}^t(\hat{\theta}_{v_{k-1}}^t)$ at trial t to a previously sampled trajectory $\hat{o}_{v,n}^1, \hat{o}_{v,n}^2, \dots, \hat{o}_{v,n}^{t-1}$. These sampled trajectories are collected throughout the interaction, such that the “trial” set is

$$D_v^{\hat{o}:t} = \{[\hat{o}_{v,n}^0, \hat{o}_{v,n}^1, \dots, \hat{o}_{v,n}^t] | \hat{o}_v^\tau \sim \hat{\pi}_{v_{k-1}}^\tau(\hat{\theta}_{v_{k-1}}^\tau), n \in [N]\} \quad (31)$$

Each element of this set of sampled sequential trajectories is then separately compressed using the gzip algorithm, generating a set of compression ratios $C^t = \{c_1^t, c_2^t, \dots, c_N^t\}$, where N is the sample size and c_n^t is the compression ratio of the n th sampled trajectory. The observed sequence is also compressed, and its compression ratio c_O^t is compared to the distribution of the sampled trajectories compression ratio set. The parameter δ is used to set the percentile at which the comparison takes place. The observed trajectory is classified as typical if c_O^t is larger or equal than the δ percentile of C^t or smaller or equal to the $1 - \delta$ percentile of C^t . In this context δ can be seen as a surrogate p-value, governing the location of the observed sequence compared to the theoretical ones. Given that the set C^t can be thought of as a sample of compression ratios, the width (uncertainty) of this surrogate “confidence-interval” is governed by the size of the sample size N . Small N may lead to high uncertainty which damages the ability of this component to properly asses if the observed compression ratio fits in the theoretic set (error reduces proportionally to $\sqrt{N}$). In this work we used $N = 200$ meaning that the error was reduced by a factor of more than 10.

The typicality-based component of the N-mechanism, denoted by $Z^1(\Theta_k, h^t, \delta)$, outputs a binary vector, per trial t . Each entry in the vector indicates whether or not the observed sequence belongs to the typical set of $\theta_v \in \Theta_k$. The algorithm is defined by the environment (opponent models).

The second component, denoted by $Z^2(\Theta_k, h^t, r^t, \omega)$, verifies the opponent type by comparing the expected cumulative reward to the observed cumulative reward. In any MARL task, agents are motivated to maximise their utility. In mixed-motive and zero-sum games, the deceiving agent increases its portion of the joint reward by reducing the victim’s reward. This behaviour will contradict the victim’s expectations to earn more (due to the assumption that it has a higher DoM level). Since gaining (subjective) utility is the victim’s ultimate aim, we can expect the victim to be particularly sensitive to behaviour that leads to deviation from the expected reward. A similar idea was proposed by Oey et al. (2023), where the authors suggested the deception detection mechanism in humans is governed by the possible payoffs. Due to the coupling between the agent’s reward and its actions, this component is based on the history-conditioned expected reward $\hat{r}_\mu^t$, by averaging the expected actions and reactions per behavioural type. Formally, the expected reward, per opponent’s behavioural type, is:

$$\hat{r}_\mu^t(\theta_{v_{k-1}}) = E_{a_\mu^t \sim \pi_{\mu_k}^t} [E_{a_v^t \sim \hat{\pi}_{v_{k-1}}^t} (u_\mu(a_\mu^t, a_v^t)) | h^{t-1}] \quad (32)$$

Similarly to the gzip typicality component, this component also utilises counterfactuals to asses whether or not the observed sequence of rewards “fits” a presumed generative model. At each trial, the component samples actions from the simulated policy of each modelled type (Eq. 12) and, using the victim’s policy, computes a vector

of sampled rewards. These rewards are appended to previously sampled rewards, yielding the expected reward set:

$$R_v^{\hat{0}:t} = \{[\hat{r}_{\mu,n}^0, \hat{r}_{\mu,n}^1, \dots, \hat{r}_{\mu,n}^t] | a_\mu^t \sim \pi_{\mu_k}^t, a_v^t \sim \hat{\pi}_{v_{k-1}}^t, n \in [N]\} \quad (33)$$

At each trial, this component verifies if the cumulative *observed* reward: $\sum_{i=1}^t r_\mu^i$ belongs to the set of *expected* cumulative rewards:

$$CR_\mu^t = \{\sum_{i=1}^t \hat{r}_{\mu,n}^i | n \in [N]\} \quad (34)$$

This is determined by a parameter $\omega \in [0, 1]$ that is used similarly to δ in the gzip components, determining the percentiles that the observed cumulative reward needs to satisfy to be counted as ‘typical’ reward. This component’s output is similar to the output of $Z^1(\Theta_k, h^t)$, namely a vector of size $|\Theta_k|$ with each entry is a binary variable, indicating if the observed reward is within acceptable bounds, implying whether the reward could have been generated by a $\theta_v \in \Theta_k$ opponent. Notably, if the expected reward is higher than the upper limit, this component is also activated, even though such an event benefits the victim. Of course, this component may be tuned to alert only when the expected reward is too low.

The two components are correlated, as μ ’s reward, monitored by Z^2 is a function of v ’s actions, which are monitored by Z^1 . We combine the output of the components using element-wise logical conjunction recursively: $f^t = f^{t-1} \wedge (Z^1(\cdot) \wedge Z^2(\cdot))$. The recursive update, similar to the Bayesian update, is inspired by the assumption that agents cannot change their type during an interaction (Gmytrasiewicz and Doshi 2005)—meaning that if a certain type is excluded by the **N**-mechanism at time t it cannot be verified in the future. The full mechanism is described in Algorithm 1. The output, a binary vector, is then one of the inputs to the **N**-policy.

Algorithm 1 **N**-mechanism

Input: $f^{t-1}, \Theta_k, h^t, r^t, \delta(t), \omega$

Output: f^t

```

1: procedure N-MECHANISM( $f^{t-1}, \Theta_k, h^t, r^t, \delta(t), \omega, N$ )
2:    $x \leftarrow Z^1(\Theta_k, h^t, \delta(t), N)$  ▷ Environment dependent
3:    $y \leftarrow Z^2(\Theta_k, h^t, r^t, \omega)$ 
4:    $f^t \leftarrow f^{t-1} \wedge (x \wedge y)$ 
5:   return  $f^t$ 
6: end procedure
7: procedure  $Z_{\text{GZIP}}^1(\Theta_k, h^t, \delta(t), N)$  ▷ From Alg. 2
8:   Compute  $Y_{\delta(t)}^t(\theta)$  ▷ From 30
9:    $x \leftarrow h^t \in Y_{\delta(t)}^t(\theta)$ 
10:  return  $x$ 
11: end procedure
12: procedure  $Z^2(\Theta_k, h^t, r^t, \omega)$ 
13:   Sample counterfactual reward set ▷ From 33
14:   Compute cumulative reward set ▷ From 34
15:   Compute cumulative empirical reward:  $CR_\mu^t = \sum_{i=1}^t r_\mu^i$ 
16:    $q_\omega, q_{1-\omega} \leftarrow$  Compute  $\omega$  and  $1 - \omega$  percentile of  $\hat{CR}_\mu^t$ 
17:    $y \leftarrow q_\omega \leq CR_\mu^t \leq q_{1-\omega}$ 
18:   return  $y$ 
19: end procedure

```

Algorithm 2 gzip-compression

Input: Θ_k, h^t, δ, N

Output: x

```

1: procedure  $Z_{\text{gzip}}^1(\Theta_k, h^t, \delta, N)$ 
2:    $c_O^t \leftarrow$  Compute compression ratio of observation from  $h^t$  using gzip
3:   for  $\theta_{v_{k-1}}^t \in \Theta_k$  do:
4:     Generate sample  $D^t$  of size  $N$  observations  $o_i^t$  from  $\hat{\pi}_{v_{k-1}}^t(\hat{\theta}_{v_{k-1}}^t)$ 
5:     Append  $D^t$  to  $D^{0:t-1}$ 
6:   end for
7:    $C^t \leftarrow$  Compute compression ratio for each sampled trajectory  $\hat{o}_{v,n}^t \in D^{0:t}$  using gzip
8:    $q_\delta, q_{1-\delta} \leftarrow$  Compute  $\delta$  and  $1 - \delta$  percentile of  $C^t$ 
9:    $x \leftarrow q_\delta \leq c_O^t \leq q_{1-\delta}$ 
10:  return  $x$ 
11: end procedure

```

3.1.2 N-Policy. The agent's behaviour is governed by its **N**-policy (Algorithm 3). This takes as its input the output of the **N**-mechanism and the beliefs about the opponent's type and generates an action.

If the opponent's behaviour passes the **N**-mechanism for at least one type, the agent uses its $\text{DoM}(k)$ policy, in this work a SoftMax policy. Here we use the IPOMCP algorithm for the Q-value computation (Hula, Montague, et al. 2015), but any planning algorithm is applicable. In the case that the opponent's behaviour triggers the **N**-mechanism, the victim's optimal policy switches to an out-of-belief (OOB) policy. While lacking the capacity to simulate the external opponent, the **N**-policy utilises the property highlighted above: the victim knows that the deceiver is of an unknown DoM level. If the deceiver's DoM level is higher than the victim's, it means that the deceiver can fully simulate the victim's behaviour. Hence, if the OOB policy derails the opponent's utility maximisation plans, the opponent will avoid it. Building on our axiomatic framework, we define an effective OOB policy as one that neutralizes the deceiving agent's utility gain (Axiom 1), thereby disincentivizing it from engaging in deceptive behaviour. The best response inevitably depends on the nature of the task. We illustrate generic OOB policies for two different payout structures: zero-sum and mixed-motive.

In zero-sum games, the Minimax algorithm (Shannon 1993) computes the best response in the presence of an unknown opponent. This principle assumes that the unknown opponent will try to act in the most harmful manner, and the agent should be defensive to avoid exploitation. While this policy is beneficial in zero-sum games it is not rational in mixed-motive games; it prevents the agent from taking advantage of the mutual dependency of the reward structure.

In repeated mixed-motive games, several policy prescriptions have been suggested as ways of deterring a deceptive opponent. An agent following the Grim trigger policy (Friedman 1971) responds to any deviation from cooperative behaviour with endless anti-cooperative behaviour, even at the risk of self-harm. While being efficient at deterring the opponent from defecting, this policy has its pitfalls. First, it might be that the opponent's defective behaviour is by accident and random (i.e., due to SoftMax policy), and so endless retaliation is misplaced. Second, if both players can communicate, then a warning shot is a sufficient signal, allowing the opponent to change their deceptive behaviour. This requires a different model, for example, the Communicative-IPOMDP (Gmytrasiewicz and Adhikari 2019). However, in this work, we illustrate how the possibility of a Grim trigger policy suffices to deter a savvy opponent from engaging in deceptive behaviour.

3.1.3 Mixed-Motive Game. Repeating the simulation with the **N**-IPOMDP framework shows how a power imbalance is diluted via the detection and retaliation of the **N**-mechanism, and via the Grim Trigger-based

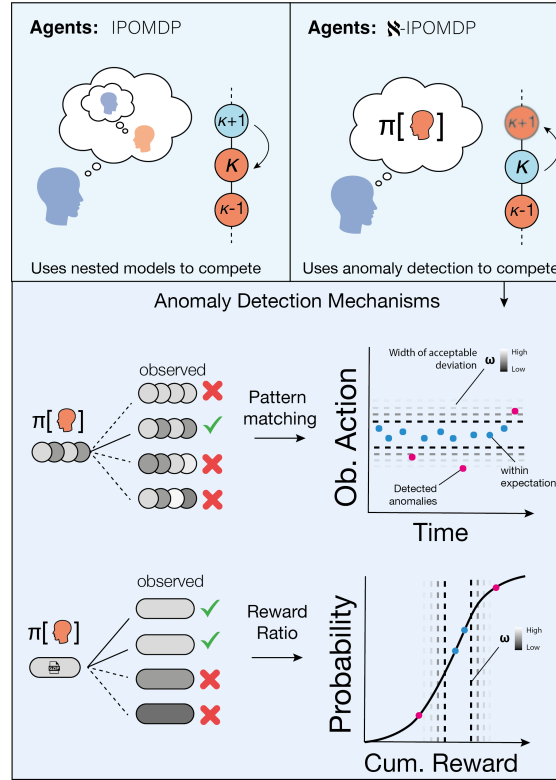


Fig. 4. **Illustration of the N-mechanism algorithm** (Top Left) In IPOMDP, $\text{DoM}(k)$ agents use their nested model to compute best-response against $\text{DoM}(k-1)$ agents, taking advantage of their superior mentalising abilities. (Top Right) N-IPOMDP augmentations allow agents to detect when their partners might be hierarchically superior, utilizing anomaly detection methods to avoid exploitation. (Bottom) The N-mechanism combines expectation-observation monitoring with typicality (policy predictions) to verify that the observed agent acts “as expected”. This can vary by agents, with individual differences dictating how much ‘evidence’ of anomalies is required before the mechanism is triggered, and also how narrow or broad the typical set is.

Algorithm 3 N-policy

Input: $b_{\mu_k}^t(\theta_{v_{k-1}}), f$
Output: a^{t+1}

```

1: procedure N-POLICY( $b_{\mu_k}^t(\theta_{v_{k-1}}), f$ )
2:   if  $f \neq 0$  then:
3:      $a^{t+1} \sim \pi_{\mu_k}^t(b_{\mu_k}^t(\theta_{v_{k-1}}))$ 
4:   else
5:      $a^{t+1}$  is sampled from OOB policy
6:   end if
7:   return  $a^{t+1}$ 
8: end procedure
  
```

N-policy. The effect of the **N**-IPOMDP is illustrated in Figure 5 depicting how the deceiver’s behaviour is affected by the **N**-mechanism and the **N**-policy (compare Figures 3(A) and 5(A)). Each component of the **N**-mechanism limits the freedom of action of the deceiver if they are to avoid the consequences of detection.

As illustrated in Figure 5, these constraints have different effects on the two deceptive DoM(1) senders in the IUG task: while both DoM(1) senders attempt to masquerade as random senders, their behaviour drastically changes compared to the IPOMDP case. First, since the expected cumulative reward component constrains them, these senders can no longer make only low offers; their offers vary to ensure that the cumulative reward is within the bounds dictated by ω . Second, the typicality monitoring (δ) forces them to diversify their offers to avoid over-efficient compression. These adapted behaviours are presented for two sets of δ, ω in Fig. 5(A,C)

Overall, our proposed mechanism reduces the income gap between the agents, as illustrated in Fig. 5(B,D). Thus it limits the ability of a deceptive, higher DoM agent to take advantage of a limited computational victim.

3.1.4 Effect of N-Mechanism Parameters. The size of the inequality reduction is a function of the **N**-mechanism parameters. Narrowing the expected reward bounds, using a small ω , forces the deceiver to make offers that are closer to the offers that the agent it is masquerading as would make, reducing the size of the set of available actions (to avoid alerting the victim). Setting larger values of δ limits the deceiver’s ability to repeat the same offer several times. The combination of the two determines the outcome of the game. However, this rigidity may harm the victim’s performance when interacting with a genuinely random agent, as could in fact be the case in the IUG task. Hence, setting these parameters requires a delicate balance between false and true negatives. To examine this, we simulated the DoM(0) **N**-IPOMDP receiver against a DoM(1) sender and a random sender over a grid of 5 different values of δ and 5 values of ω . Each design was simulated 50 times with different random seeds.

We first discuss the effect of the parameters on the receiver’s reward, as the goal of the **N**-IPOMDP is to improve the victim’s performance via deterrence. The average reward of the receiver is depicted in Fig. 6. We begin with the results against a random sender (left-most panel)—here we clearly see the detrimental results of high levels of ω and δ —aimed at verifying “genuine” random behaviour. As shown in Fig. 6 (left column), there is a partial trade-off between δ and ω when interacting with a random sender. The higher these parameters, the more sensitive the **N**-mechanism is to deviations from “truly” random behaviour. In turn, this causes the **N**-mechanism to flag a genuine random sender as being non-random, thus activating the **N**-policy and reducing the receiver’s average reward per trial compared to the “naïve” case.

The effect of the parameters is also substantial when engaged with the DoM(1) sender. Here we compare the average reward of the **N**-IPOMDP receiver to the “sucker payoff” (0.1). While the framework as a whole limits the DoM(1) sender’s ability to deceive the DoM(0) receiver, certain settings allow the DoM(1) larger wiggle room. As alluded to above, while the overall effect of the **N**-mechanism is a reduction in the reward ratio, the causal effect differs between the threshold senders. The high DoM(1) threshold sender ($\psi = 0.5$) only engages with the DoM(0) **N**-agent when the parameters are low, as any other setting coerces it to make offers that are detrimental to it. This avoidance of engagement illustrates how the **N**-IPOMDP framework serves as a deterrence mechanism—the sender is aware of the “commitment” of the receiver to “cut off its nose to spite both faces”, i.e., the mechanism serves as a credible threat, dissuading the high threshold sender from engaging with it.

On the other hand, in order to maximize reward, the low threshold ($\psi = 0.1$) DoM(1) sender alters its behaviour (compared to the deceptive baseline) to gain reward, where possible. It continues to masquerade as being random, but its policy adheres to the statistical constraints imposed by the **N**-mechanism, until it is compelled to make offers that do not benefit it, in which case its policy triggers the **N**-mechanism, ending the interaction. This effect is evident in Fig. 6 (middle column). Some combinations of (ω, δ) do not harm the receiver’s outcome substantially compared to the non-regulated game, but others reduce the receiver’s reward by more than 50%. This phenomenon illustrates Goodhart’s law—the deceiver uses its nested model to estimate, using simulation,

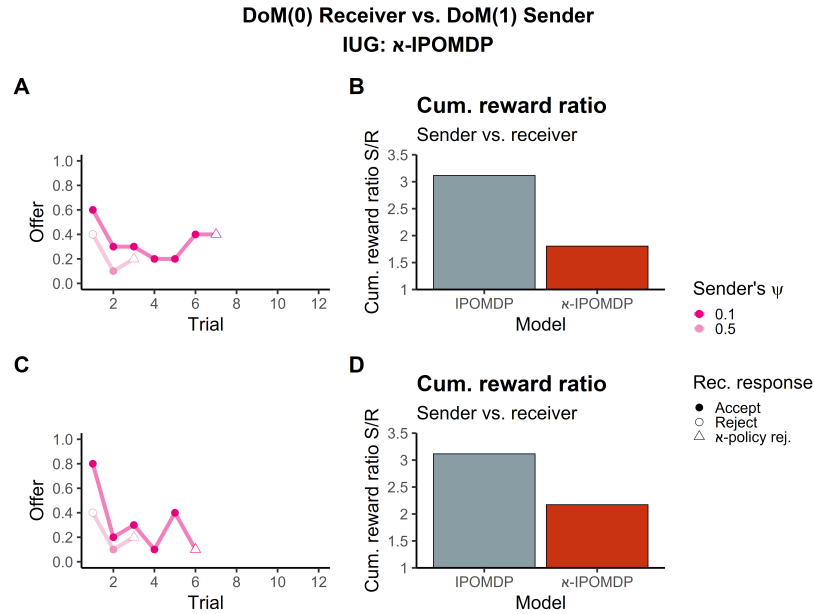


Fig. 5. **Mitigation of deception in IUG with κ -IPOMDP**: Points represent offers from the sender to the receiver across trials, biased by the sender's threshold. Points are shaded white if the receiver rejects the offer, triangular points indicate that the rejection is caused by the κ -mechanism, effectively terminating the interaction. Lines and points are visible while the κ -mechanism is off.

Top row $\delta = 0.1$, $\omega = 0.3$ — **(A)** Notably, both DoM(1) senders masquerade as being random. However, their ability to execute the “random” behaviour ruse is limited by both κ -mechanism components. First, the cumulative reward has to satisfy the off-policy counterfactual reward component. Next, the variability of the offers is higher than in the IPOMDP case, respecting the typicality component. Ultimately, the deceiver's policy triggers the average reward monitoring (Z^2) component, as the observed reward is lower than expected (marked by the truncation of the line and points). **(B)** Cumulative reward ratio for the sender vs. the receiver. The κ -IPOMDP reduces the cumulative reward ratio (sender/receiver) by more than 40%.

Bottom row $\delta = 0.3$, $\omega = 0.3$ — **(C)** When it is constrained by narrower strong typicality set (Z^1) bounds, the DoM(1) with low threshold terminates the interaction faster than before, triggering this component after 6 trials, while the high threshold sender acts similarly. **(D)** Even when the interaction is shorter, the reward ratio is still reduced compared to the case of the conventional IPOMDP.

how to avoid detection, at a cost. While the κ -IPOMDP framework fails to coerce the deceiver into disclosing its true type fully (and in fact it encourages it to learn a new deceptive move) it succeeds in mitigating the effect of deception, namely decreasing the gain from deceptive behaviour and reducing the reward gap.

κ -Mechanism Components Effect. As illustrated in Fig. 5, the κ -mechanism disturbs the deceptive behaviour of the DoM(1) sender, causing it to act in a more random manner, avoiding detection as long as possible. This modified deceptive behaviour is still picked by the κ -mechanism, but which of the two components is more likely to detect the mismatch between the expected and observed behaviour is triggered first? We answer this question by plotting the probability that the κ -mechanism is triggered first by the strong-typicality component (Z^1) by sender threshold, depicted in Fig. 7. In this figure we depict the probability that the κ -mechanism is activated by

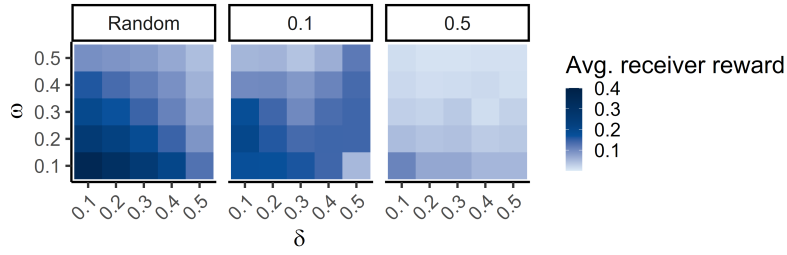


Fig. 6. **Effect of N-mechanism parameters on IUG outcome for DoM(0) receiver interacting with random and DoM(1) sender:** Measuring the receiver’s average reward (averaged across 20 simulations) in the IUG with different, fixed, values of ω (y-axis) and δ (x-axis) reveals how the parameters of the N-mechanism affect the interaction. When engaged with a random sender (Left), tight constraints lead to an overactive N-mechanism that nullifies the interaction. This is because the random sender does not alter its behaviour in response to rejections by the receiver. In this case, the DoM(0) receiver needs to be very flexible in its N-mechanism to collect a reward. As evident in the case of lower values of δ and ω - even widely tuned mechanism leads to a decrease in the outcome, from expected 0.5 to about 0.4. When interacting against the threshold DoM(1) senders different parameter combinations yield rewards for the receiver. In the case of the lower threshold sender, most combinations improve the receiver’s performance relative to the “sucker payoff” of 0.1—in particular the combination of wide strong typicality set (Z^1 component, low δ) and tighter average reward bounds (Z^2 component, small ω) doubles the “sucker payoff” for the receiver. On the other hand, limited by the N-mechanism, the high threshold sender activates the punitive policy fast, rendering the average reward to lower than the “sucker payoff”, but crucially — lowers substantially the sender’s reward too. Only when both parameters are set to 0.1 the receiver’s average payout outperforms the “sucker payoff”.

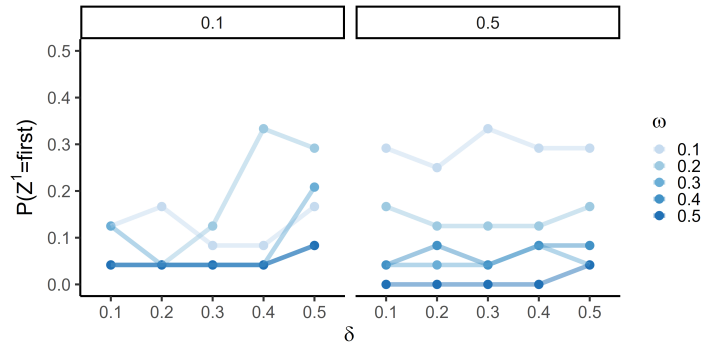


Fig. 7. **Probability of activation by Z^1 .** The points show the probability that the N-Mechanism is activated by the strong-typicality component prior to being activated by the reward monitoring component (Z^2). Lines indicate the size of the ω and each column represents a different sender threshold. This plot show that the N-mechanism is more likely to be activated by the reward monitoring. However, while high values of ω (narrow reward bounds) suppress the Z^1 component, when these bounds are quite wide, the strong typicality component is more likely to activated ($\omega \in [0.1, 0.2]$). Moreover, the Z^1 mechanism is more likely to be activated when interacting with the higher threshold sender

Z^1 before being triggered by Z^2 . This plot shows that overall the mechanism is more susceptible to deviation from expected reward rather than to deviation from expected compression ratio, but this effect differs by the sender’s threshold and associated behaviour. We included an analysis by trial in Appendix 9.

N-Mechanism False Positives. While effective in deterring higher DoM agents from engaging in protracted deceptive behaviour, the **N**-mechanism may backfire when interacting with a genuine agent—in this case, the random sender (Alon, Schulz, Dayan, et al. 2023). This false detection depends on the parameters of both components. We begin with an overall analysis of false detection illustrated in Fig. 8. This plot depicts the probability that the mechanism will falsely activate when engaging with a random sender as a function of the interaction duration. As evident the tighter the bounds (higher parameters) the more likely it is that the mechanism will mistakenly flag the sender as non-random, terminating the interaction. This plot shows evidence that the **N**-mechanism is overly aggressive. Future work may resolve this issue by balancing false positives and negatives. This finding better explains the poor results depicted in Fig. 6(Left panel). Crucially, as in the case of the truly non-random sender the mechanism is more sensitive to deviations from strong typicality in early trials and is more vulnerable to misclassifications by the average reward monitoring in later trials.

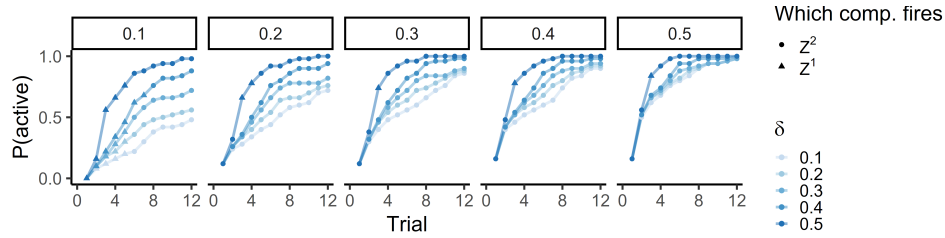


Fig. 8. **N-mechanism false positive probability.** The **N**-mechanism misdetects a random sender as non-random, causing the **N**-IPOMDP agent to wrongly terminate the interaction. This plot depicts this false-positive probability (y-axis) as a function of the trial number (x-axis) and the **N**-mechanism components parameters— ω (columns) and δ (colours). When the parameters are low enough, the probability of false-positive is lower than 0.5 (for the case of $\omega = 0.1, \delta = 0.1$), but spikes up to almost 1 when the parameters are high. This false identification is caused by an overly sensitive and narrow window of acceptable deviation from expected outcomes.

Thus, in the spirit of a “no free lunch” theorem, our results show that the best set of parameters against the low threshold DoM(1) sender is not best against the high threshold DoM(1) sender nor against a random one. However, on average, certain sets of parameters yield the desired outcome—mitigating deceptive behaviour.

4 Discussion

We imbued agents with the ability to assess whether they are being deceived, without (at least fully) having to conceptualise how. To achieve this we augmented Bayesian inference within the **N**-IPOMDP mechanism with environmental and anomaly detection mechanisms. The **N**-policy of these agents is the ability to infer that they are facing an agent outside their world model that threatens to harm them. The net result is a more equitable outcome. This framework offers a new conceivable algorithm to tackle limited resources in cognitive hierarchies instigated by the IPOMDP. As well as providing healthy strategic detection, perturbing our framework offers insight into mechanisms underlying overly sensitive anomaly detection, evidenced in paranoia, psychosis, and conspiracy theory.

We tested the **N** mechanism in representative mixed-motive and zero-sum Bayesian repeated games. We show that the **N** mechanism can protect less sophisticated agents against greedy, clever deceivers who utilize nested models to coerce their partners. Such a pretence deployed by a sophisticated sender depends on atypical sequences of actions. Usually, detecting this is outside the scope of a lower DoM agent, but the **N**-mechanism allows a measure of protection. Of course, the higher DoM agent can, at least if well-calibrated with its simpler partner, predict exactly when the **N**-mechanism will fire, and take tailored offensive measures. However, the net

effect in both games shows that their hands might be sufficiently tied to make the outcome fairer. Even with the increased deception complexity available to higher DoM agents ($k \geq 2$) (Alon, Schulz, Rosenschein, et al. 2023b), we expect that agents endowed with the **N**-mechanism, will cope with deceptive ploys in a similar manner to the blueprints presented in this work. We leave this evaluation for future work.

The two defining points for the **N**-IPOMDP framework are augmented inference (the **N**-mechanism), and an appropriate response (the **N**-policy). For the inference, there could be, as we present here, exquisitely tailored parameters that perform most competently in a given interaction; however, generalisation is hard. Future work may explore how to incorporate learning into parameter tuning, making the parameter setting an adaptive process rather than a fixed one.

Our framework is built on prior work formalising irritation in the context of the multi-round trust task (Hula, Vilares, et al. 2018), in which total non-cooperation is a consequence of high irritation. This is a blunt, and deliberately self-destructive instrument to be a credible and thus an effective deterrent (McNamara and Houston 2002). One alternative might be a decision that it is worth investing more cognitive effort, such that the victim increases their DoM (Alaoui and Penta 2016; Yu et al. 2021), although this could easily become a cognitively expensive arms race (S. Sarkadi 2023). Instead, we offer a less computationally intensive manner to allow agents with limited resources to mitigate manipulation.

One limitation is that our model cannot reason about the intentions and consequent plans of the deceiver, which may be crucial to facilitate opponent learning (as in the work of Yu et al. (2021), where agents can learn how to adapt their recursive level via learning). A DoM(k) agent would benefit from such an ability (say via self-play) in repeated interactions, or from the ability to engage in epistemic planning (Belle et al. 2023). However, a savvy opponent, aware of this learning and planning capacity, can still manipulate the learning process to its benefit (Foerster et al. 2018), although such a setting requires substantially more computation due to the recursive nature of the problem.

Another issue, keeping with its roots in competitive economics, is our focus on how lower DoM agents might be exploited. One could also imagine the case that the higher DoM agent *exceeds* expectations by sharing more than the lower DoM agent expects. Although this problem may be solved with an editing of the **N**-policy, it could also be a sign that the higher DoM has a social orientation ‘baked’ into its policy to a greater extent than expected. In this case, the lower DoM agent might want to have the capacity to compensate for the over-fair actions and not break down. Naturally, these mechanisms are prone to excess manipulation, and so would need careful monitoring. On the other hand, in continuation of the Machiavellian Theory-of-Mind origin theory by Byrne and Whiten (1988), such a beneficiary behaviour may result from an elaborate ruse, designed to fool the victim until the bitter end (like the evil witch in Hansel and Gretel), and hence the **N**-IPOMDP should also account for last minute betrayal by acting defensively.

Overreacting may also arise when the mismatch stems not from strategic manipulation but from simple model error (Stahl II and Wilson 1994; Wright and Leyton-Brown 2010) or a discrepancy between the actual and assumed prior distributions over components of the opponent, i.e., misalignment. This issue gives rise to several related problems. First, as illustrated, with the spirit of the no free lunch theorem, the parameters of the **N**-policy need to balance sensitivity and specificity. Hence, the mechanism might either reduce false alarms at the expense of missing true deception or might be overactive and cause the victim to misclassify truly benign behaviour. The latter could then result in paranoid-like behaviour, offering an alternative explanation to paranoia through over-mentalising (Alon, Schulz, Dayan, et al. 2023). Erroneous anomaly detection has long been theorised as a potential key precursor to paranoid, delusion-like beliefs (Howes and Murray (2014)), although to date, few theories have been able to formally translate into appropriate contexts observed in the clinic, which invariably involve social and intentional concerns (Barnby et al. 2023). This theory moves toward plurality in providing sufficient and necessary computational mechanisms of paranoid-like inference. A second problem is that the model currently assumes k -level reasoning. This means that the **N**-mechanism is activated by agents with lower

than $(k - 1)$ DoM level. However, a DoM(k) is fully capable of modelling these agents, as they are part of its nested opponent models. Thus, following the ideas suggested by (Camerer et al. 2004), future work may extend the opponent set to include all DoM levels up to k . Overreaction may arise from an innocent mistake (for example due to high SoftMax temperature) rather than a deceptive move. To mitigate the severe response, it might be beneficial to incorporate a confidence level to the $\mathbf{N}$ -mechanism’s output, turning it from a binary vector into a probabilistic estimation of the “Goodness of fit” of the observed behaviour to the modelled one. Such relaxation is used in Psychology and language (Collins and Hahn (2020)) to allow agents to quantify the reliability of an utterance.

Lastly, an issue with inference arises from our assumption that deceptive behaviour can be detected by measuring deviation from typical behaviour. This assumption is challenged in two ways. First, a well-disguised opponent may imitate the behaviour in such a way that is not deviating enough from the assumed typical behaviour or perfectly masquerading in a way that is undetectable by our proposed mechanism. Second, if agents’ behaviour becomes more erratic (associated with a lower inverse temperature in the SoftMax policy), the ability to detect and classify both “expected” and “unexpected” behaviour will be jeopardized (Alon, Schulz, Bell, et al. 2024). In this case, the anomaly detection should take into consideration the increased probability of off-path behaviour and adapt accordingly. We leave room for future work to solve these issues.

Opponent verification is relevant to cybersecurity (Obaidat et al. 2019), where legitimate users need to be verified and malevolent ones blocked. However, savvy hackers learn to avoid certain anomalies while still exploiting the randomness of human behaviour. To balance effectively between defence and freedom of use, these systems need to probe the user actively to confirm the user’s identity. Our model proposes one such solution but lacks the active learning component, which future work may incorporate. Another important ramification of our work is in the application to AI safety and alignment. Recent interest has flourished concerning the emergence of ToM-like ability in LLMs (Kosinski 2023; Sap et al. 2022; Ullman and Bass 2024) and simultaneously the risk this brings with aligning LLM goals with human goals. As argued in our paper and hinted at by Van Ditmarsch et al. (2020), if these models possess high levels of ToM ability, it is not unlikely that they will use it to manipulate human users if it benefits them Sabour et al. (2025). One way of mitigating such a behaviour is via implementation of the $\mathbf{N}$ -IPOMDP for this problem. Our model proposes blueprints for intention verification to avoid model tampering. If an LLM is maliciously used to affect the user’s beliefs, an $\mathbf{N}$ -like model, applied on the output of the models, may detect a mismatch between the model’s output and the expected one for a given context. Coupled with a proper “punishment” mechanism (like blocking users from using the model) it may deter perpetrators from engaging in such social engineering endeavours.

The $\mathbf{N}$ -IPOMDP framework represents a significant advance in protecting agents from deception and manipulation in adversarial environments. By augmenting Bayesian inference with the $\mathbf{N}$ -mechanism we provide a robust method for less sophisticated agents to detect and respond to potentially harmful behaviour from higher-ToM deceivers, a possibility previously out of reach in a typical k -level hierarchy. The framework’s effectiveness underscores its potential for broader applications, including cybersecurity and large language models (LLMs). Future work must address model generalisation, cognitive cost management, and the integration of active learning components to further enhance the framework’s robustness and adaptability. As we continue to explore and refine these mechanisms, the potential for creating fairer and more secure interactive systems becomes increasingly attainable.

References

- S. Adhikari and P. J. Gmytrasiewicz. 2021. “Telling Friend from Foe-Towards a Bayesian Approach to Sincerity and Deception.” In: *TRUST@AAMAS*.
- L. Alaoui and A. Penta. Oct. 2016. “Endogenous Depth of Reasoning.” *The Review of Economic Studies*, 83, 4, (Oct. 2016), 1297–1333. doi:10.1093/restud/rdv052.

- N. Alon, L. Schulz, V. Bell, M. Moutoussis, P. Dayan, and J. M. Barnby. 2024. "(Mal) adaptive mentalizing in the cognitive hierarchy, and its link to paranoia." *Computational Psychiatry*, 8, 1, 159.
- N. Alon, L. Schulz, P. Dayan, and J. M. Barnby. 2023. "Between prudence and paranoia: Theory of Mind gone right, and wrong." In: *First Workshop on Theory of Mind in Communicating Agents*. <https://openreview.net/forum?id=gB9zrEjhZD>.
- N. Alon, L. Schulz, J. S. Rosenschein, and P. Dayan. 2023a. "A (Dis-) information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind." *Open Mind*, 7, 608–624.
- N. Alon, L. Schulz, J. S. Rosenschein, and P. Dayan. Aug. 2023b. "A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind." *Open Mind*, 7, (Aug. 2023), 608–624. doi:10.1162/opmi_a_00097.
- J. M. Barnby, P. Dayan, and V. Bell. 2023. "Formalising social representation to explain psychiatric symptoms." *Trends in Cognitive Sciences*.
- V. Belle, T. Bolander, A. Herzig, and B. Nebel. 2023. *Epistemic planning: Perspectives on the special issue*. (2023).
- P. G. Bennett. Dec. 1980. "Hypergames: Developing a model of conflict." *Futures*, 12, 6, (Dec. 1980), 489–507. doi:10.1016/0016-3287(80)90005-1.
- A. Blum and Y. Mansour. Sept. 2007. "Learning, Regret Minimization, and Equilibria." en. In: *Algorithmic Game Theory*. (1st ed.). Ed. by N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. Cambridge University Press, (Sept. 2007), 79–102. ISBN: 978-0-521-87282-9 978-0-511-80048-1. doi:10.1017/CBO9780511800481.006.
- N. Buergi, G. Aydogan, A. Kononov, and C. Ruff. Feb. 2024. "A neural fingerprint of adaptive mentalization." en-us, (Feb. 2024). Publisher: OSF. doi:10.31234/osf.io/npqy8.
- R. W. Byrne and A. Whiten, (Eds.) . July 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press, Oxford, New York, (July 1988). ISBN: 978-0-19-852175-4.
- C. F. Camerer. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.
- C. F. Camerer, T.-H. Ho, and J.-K. Chong. Aug. 2004. "A Cognitive Hierarchy Model of Games*." *The Quarterly Journal of Economics*, 119, 3, (Aug. 2004), 861–898. doi:10.1162/0033553041502225.
- M. S.-Y. Chwe. 2013. *Rational ritual: Culture, coordination, and common knowledge*. Princeton University Press.
- P. J. Collins and U. Hahn. 2020. "We might be wrong, but we think that hedging doesn't protect your reputation." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46, 7, 1328–1348. Place: US Publisher: American Psychological Association. doi:10.1037/xlm000786.
- F. Cuzzolin, A. Morelli, B. Cirstea, and B. J. Sahakian. 2020. "Knowing me, knowing you: theory of mind in AI." *Psychological medicine*, 50, 7, 1057–1061.
- P. Deutsch. 1996. *GZIP file format specification version 4.3*. Tech. rep.
- M. Devaine, G. Hollard, and J. Daunizeau. Dec. 2014. "The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn?" en. *PLOS Computational Biology*, 10, 12, (Dec. 2014), e1003992. Publisher: Public Library of Science. doi:10.1371/journal.pcbi.1003992.
- M. Eger and C. Martens. 2017. "Practical Specification of Belief Manipulation in Games." en. *Proceedings of the AAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13, 1, 30–36. Number: 1. doi:10.1609/aiide.v13i1.12921.
- S. Evans, E. Eiland, S. Markham, J. Impson, and A. Laco. 2007. "MDLcompress for Intrusion Detection: Signature Inference and Masquerade Attack." In: *MILCOM 2007 - IEEE Military Communications Conference*, 1–7. doi:10.1109/MILCOM.2007.4455304.
- J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. Sept. 2018. "Learning with Opponent-Learning Awareness." en. *arXiv:1709.04326 [cs]*, (Sept. 2018). arXiv: 1709.04326. Retrieved May 13, 2020 from <http://arxiv.org/abs/1709.04326>.
- M. C. Frank and N. D. Goodman. May 2012. "Predicting Pragmatic Reasoning in Language Games." en. *Science*, 336, 6084, (May 2012), 998–998. doi:10.1126/science.1218633.
- J. W. Friedman. 1971. "A non-cooperative equilibrium for supergames." *The Review of Economic Studies*, 38, 1, 1–12.
- C. D. Frith and U. Frith. 2021. "Mapping mentalising in the brain." In: *The neural basis of mentalizing*. Springer, 17–45.
- D. Fudenberg and J. Tirole. Apr. 1991. "Perfect Bayesian equilibrium and sequential equilibrium." *Journal of Economic Theory*, 53, 2, (Apr. 1991), 236–260. doi:10.1016/0022-0531(91)90155-W.
- P. J. Gmytrasiewicz and P. Doshi. July 2005. "A Framework for Sequential Planning in Multi-Agent Settings." en. *Journal of Artificial Intelligence Research*, 24, (July 2005), 49–79. doi:10.1613/jair.1579.
- P. J. Gmytrasiewicz and S. Adhikari. 2019. "Optimal Sequential Planning for Communicative Actions: A Bayesian Approach." In: *AAMAS*. Vol. 19, 1985–1987.
- N. D. Goodman and M. C. Frank. Nov. 2016. "Pragmatic Language Interpretation as Probabilistic Inference." *Trends in Cognitive Sciences*, 20, 11, (Nov. 2016), 818–829. doi:10.1016/j.tics.2016.08.005.
- J. C. Harsanyi. 1968. "Games with Incomplete Information Played by "Bayesian" Players, I-III. Part III. The Basic Probability Distribution of the Game." *Management Science*, 14, 7, 486–502. Retrieved Dec. 11, 2023 from <http://www.jstor.org/stable/2628894>.
- O. D. Howes and R. M. Murray. 2014. "Schizophrenia: an integrated sociodevelopmental-cognitive model." *The Lancet*, 383, 9929, 1677–1687.
- H. Hu, A. Lerer, B. Cui, D. Wu, L. Pineda, N. Brown, and J. Foerster. Aug. 2021. "Off-Belief Learning." *arXiv:2103.04000 [cs]*, (Aug. 2021). arXiv: 2103.04000. Retrieved Apr. 10, 2022 from <http://arxiv.org/abs/2103.04000>.

- A. Hula, P. R. Montague, and P. Dayan. June 2015. "Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange." en. *PLOS Computational Biology*, 11, 6, (June 2015), e1004254. Publisher: Public Library of Science. doi:[10.1371/journal.pcbi.1004254](https://doi.org/10.1371/journal.pcbi.1004254).
- A. Hula, I. Vilares, T. Lohrenz, P. Dayan, and P. R. Montague. Feb. 2018. "A model of risk and mental state shifts during social interaction." en. *PLOS Computational Biology*, 14, 2, (Feb. 2018), e1005935. Publisher: Public Library of Science. doi:[10.1371/journal.pcbi.1005935](https://doi.org/10.1371/journal.pcbi.1005935).
- N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. A. Ortega, D. J. Strouse, J. Z. Leibo, and N. de Freitas. June 2019. "Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning." en. *arXiv:1810.08647 [cs, stat]*, (June 2019). arXiv: 1810.08647. Retrieved Dec. 7, 2020 from <http://arxiv.org/abs/1810.08647>.
- P. Jin, K. Keutzer, and S. Levine. Oct. 2018. *Regret Minimization for Partially Observable Deep Reinforcement Learning*. en. arXiv:1710.11424 [cs]. (Oct. 2018). Retrieved July 1, 2024 from <http://arxiv.org/abs/1710.11424>.
- L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. May 1998. "Planning and acting in partially observable stochastic domains." en. *Artificial Intelligence*, 101, 1, (May 1998), 99–134. doi:[10.1016/S0004-3702\(98\)00023-X](https://doi.org/10.1016/S0004-3702(98)00023-X).
- D.-K. Kim, M. Riemer, M. Liu, J. Foerster, M. Everett, C. Sun, G. Tesauro, and J. P. How. 2022. "Influencing long-term behavior in multiagent reinforcement learning." *Advances in Neural Information Processing Systems*, 35, 18808–18821.
- P. W. Koh and P. Liang. 2017. "Understanding black-box predictions via influence functions." In: *International conference on machine learning*. PMLR, 1885–1894.
- C. Kopp, K. B. Korb, and B. I. Mills. Nov. 2018. "Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to "fake news"." en. *PLOS ONE*, 13, 11, (Nov. 2018), e0207383. Publisher: Public Library of Science. doi:[10.1371/journal.pone.0207383](https://doi.org/10.1371/journal.pone.0207383).
- M. Kosinski. Feb. 2023. *Theory of Mind May Have Spontaneously Emerged in Large Language Models*. arXiv:2302.02083 [cs]. (Feb. 2023). doi:[10.48550/arXiv.2302.02083](https://doi.org/10.48550/arXiv.2302.02083).
- N. Le Guillarme, A.-I. Mouaddib, S. Gatepaille, and A. Bellenger. Nov. 2016. "Adversarial Intention Recognition as Inverse Game-Theoretic Planning for Threat Assessment." In: *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. ISSN: 2375-0197. (Nov. 2016), 698–705. doi:[10.1109/ICTAI.2016.0111](https://doi.org/10.1109/ICTAI.2016.0111).
- B. Liebal, D. Roth, N. Shah, and V. Srikumar. 2007. *Proactive Detection of Insider Attacks*. Tech. rep. UIUCDCS-R-2007-2879. University of Illinois.
- M. L. Littman et al.. 2001. "Friend-or-foe Q-learning in general-sum games." In: *ICML 2001*. Vol. 1, 322–328.
- P. Maguire, P. Moser, R. Maguire, and M. T. Keane. 2019. "Seeing patterns in randomness: A computational model of surprise." *Topics in cognitive science*, 11, 1, 103–118.
- P. Masters and S. Sardina. 2017. "Deceptive Path-Planning," 4368–4375. Retrieved Aug. 10, 2022 from <https://www.ijcai.org/proceedings/2017/4368>.
- P. Masters, W. Smith, L. Sonenberg, and M. Kirley. 2021. "Characterising deception in AI: A survey." In: *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*. Springer, 3–16.
- J. M. McNamara and A. I. Houston. 2002. "Credible threats and promises." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357, 1427, 1607–1616.
- R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht. 2022. "A survey of ad hoc teamwork research." In: *European conference on multi-agent systems*. Springer, 275–293.
- S. Na et al.. Oct. 2021. "Humans use forward thinking to exploit social controllability." *eLife*, 10, (Oct. 2021), e64983. Ed. by C. Hartley, C. Büchel, C. Hartley, and R. Ligneul. Publisher: eLife Sciences Publications, Ltd. doi:[10.7554/eLife.64983](https://doi.org/10.7554/eLife.64983).
- J. H. Nachbar and W. R. Zame. 1996. "Non-computable strategies and discounted repeated games." *Economic theory*, 8, 103–122.
- M. S. Obaidat, I. Traore, and I. Woungang. 2019. *Biometric-based physical and cybersecurity systems*. Springer.
- L. A. Oey, A. Schachner, and E. Vul. 2023. "Designing and detecting lies by reasoning about other agents." *Journal of Experimental Psychology: General*, 152, 2, 346.
- E. Pacuit and O. Roy. 2017. "Epistemic Foundations of Game Theory." In: *The Stanford Encyclopedia of Philosophy*. (Summer 2017 ed.). Ed. by E. N. Zalta. Metaphysics Research Lab, Stanford University. Retrieved June 28, 2023 from <https://plato.stanford.edu/archives/sum2017/entries/epistemic-game/>.
- J. Palomäki, J. Yan, and M. Laakasuo. Aug. 2016. "Machiavelli as a poker mate — A naturalistic behavioural study on strategic deception." *Personality and Individual Differences*, 98, (Aug. 2016), 266–271. doi:[10.1016/j.paid.2016.03.089](https://doi.org/10.1016/j.paid.2016.03.089).
- G. Pannell and H. Ashman. 2010. "User Modelling for Exclusion and Anomaly Detection: A Behavioural Intrusion Detection System." en. In: *User Modeling, Adaptation, and Personalization*. Ed. by P. De Bra, A. Kobsa, and D. Chin. Springer, Berlin, Heidelberg, 207–218. ISBN: 978-3-642-13470-8. doi:[10.1007/978-3-642-13470-8_20](https://doi.org/10.1007/978-3-642-13470-8_20).
- J. Peng, K.-K. R. Choo, and H. Ashman. Sept. 2016. "User profiling in intrusion detection: A review." *Journal of Network and Computer Applications*, 72, (Sept. 2016), 14–27. doi:[10.1016/j.jnca.2016.06.012](https://doi.org/10.1016/j.jnca.2016.06.012).
- N. Piazza, V. Behzadan, and S. Sarkadi. Nov. 2023. *Limitations of Theory of Mind Defenses against Deception in Multi-Agent Systems*. en. preprint. In Review, (Nov. 2023). doi:[10.21203/rs.3.rs-3518720/v1](https://doi.org/10.21203/rs.3.rs-3518720/v1).

- D. Premack and G. Woodruff. Dec. 1978. "Does the chimpanzee have a theory of mind?" en. *Behavioral and Brain Sciences*, 1, 4, (Dec. 1978), 515–526. Publisher: Cambridge University Press. doi:[10.1017/S0140525X00076512](https://doi.org/10.1017/S0140525X00076512).
- D. Ramachandran and E. Amir. 2007. "Bayesian inverse reinforcement learning." In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. Morgan Kaufmann Publishers Inc., Hyderabad, India, 2586–2591.
- M. Ramírez and H. Geffner. 2011. "Goal recognition over POMDPs: inferring the intention of a POMDP agent." In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three (IJCAI'11)*. AAAI Press, Barcelona, Catalonia, Spain, 2009–2014. ISBN: 9781577355151.
- I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach. 2021. "Adversarial machine learning attacks and defense methods in the cyber security domain." *ACM Computing Surveys (CSUR)*, 54, 5, 1–36.
- T. Rusch, S. Steininger-Kumar, P. Doshi, M. Spezio, and J. Gläscher. Sept. 2020. "Theory of mind and decision science: Towards a typology of tasks and computational models." en. *Neuropsychologia*, 146, (Sept. 2020), 107488. doi:[10.1016/j.neuropsychologia.2020.107488](https://doi.org/10.1016/j.neuropsychologia.2020.107488).
- S. J. Russell. 2010. *Artificial intelligence: A modern approach*. Pearson Education, Inc.
- S. Sabour et al.. 2025. "Human Decision-making is Susceptible to AI-driven Manipulation." *ArXiv*, abs/2502.07663. <https://api.semanticscholar.org/CorpusID:276258858>.
- M. B. Salem, S. Hershkop, and S. J. Stolfo. 2008. "A Survey of Insider Attack Detection Research." en. In: *Insider Attack and Cyber Security: Beyond the Hacker*. Advances in Information Security. Ed. by S. J. Stolfo, S. M. Bellovin, A. D. Keromytis, S. Hershkop, S. W. Smith, and S. Sinclair. Springer US, Boston, MA, 69–90. ISBN: 978-0-387-77322-3. doi:[10.1007/978-0-387-77322-3_5](https://doi.org/10.1007/978-0-387-77322-3_5).
- M. Sap, R. LeBras, D. Fried, and Y. Choi. Oct. 2022. *Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs*. arXiv:2210.13312 [cs]. (Oct. 2022). doi:[10.48550/arXiv.2210.13312](https://doi.org/10.48550/arXiv.2210.13312).
- S. Sarkadi. 2023. "An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies." In: *4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems ACSOS 2023*. IEEE Computer Society.
- S. Sarkadi. 2021. "Deception." phd. University of London, King's College (United Kingdom). AAI28931207.
- Ş. Sarkadi, A. R. Panisson, R. H. Bordini, P. McBurney, S. Parsons, and M. Chapman. Jan. 2019. "Modelling deception using theory of mind in multi-agent systems." en. *AI Communications*, 32, 4, (Jan. 2019), 287–302. Publisher: IOS Press. doi:[10.3233/AIC-190615](https://doi.org/10.3233/AIC-190615).
- Ş. Sarkadi, A. Rutherford, P. McBurney, S. Parsons, and I. Rahwan. Sept. 2021. "The evolution of deception." *Royal Society Open Science*, 8, 9, (Sept. 2021), 201032. Publisher: Royal Society. doi:[10.1098/rsos.201032](https://doi.org/10.1098/rsos.201032).
- Y. Savas, C. K. Verginis, and U. Topcu. June 2022. "Deceptive Decision-Making under Uncertainty." en. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 5, (June 2022), 5332–5340. Number: 5. doi:[10.1609/aaai.v36i5.20470](https://doi.org/10.1609/aaai.v36i5.20470).
- C. E. Shannon. 1993. "Programming a computer for playing chess." In: *first presented at the National IRE Convention, March 9, 1949, and also in Claude Elwood Shannon Collected Papers*. IEEE Press, 637–656.
- H. H. da Silva, M. Rocha, G. Trajano, A. S. Morales, S. Sarkadi, and A. R. Panisson. 2024. "Distributed Theory of Mind in Multi-Agent Systems." In: *16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*. SciTePress.
- S. Siu. Dec. 2022. *Towards automating Codenames spymasters with deep reinforcement learning*. arXiv:2212.14104 [cs]. (Dec. 2022). doi:[10.48550/arXiv.2212.14104](https://doi.org/10.48550/arXiv.2212.14104).
- D. O. Stahl II and P. W. Wilson. 1994. "Experimental evidence on players' models of other players." *Journal of economic behavior & organization*, 25, 3, 309–327.
- P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein. 2010. "Ad hoc autonomous agent teams: Collaboration without pre-coordination." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 1. Vol. 24, 1504–1509.
- R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: an introduction*. en. (Second edition ed.). Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts. ISBN: 978-0-262-03924-6.
- R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. 1999. "Policy Gradient Methods for Reinforcement Learning with Function Approximation." In: *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. event-place: Denver, CO. MIT Press, Cambridge, MA, USA, 1057–1063. Retrieved Dec. 17, 2019 from <http://dl.acm.org/citation.cfm?id=3009657.3009806>.
- T. Ullman and I. Bass. May 2024. *The Detection of Automatic Behavior in Other People*. en-us. (May 2024). doi:[10.31234/osf.io/8r4yf](https://doi.org/10.31234/osf.io/8r4yf).
- H. Van Ditmarsch, P. Hendriks, and R. Verbrugge. 2020. "Editors' review and introduction: Lying in logic, language, and cognition." *Topics in cognitive science*, 12, 2, 466–484.
- N. Wang, J. Han, and J. Fang. June 2012. "An Anomaly Detection Algorithm Based on Lossless Compression." In: *2012 IEEE Seventh International Conference on Networking, Architecture, and Storage*. (June 2012), 31–38. doi:[10.1109/NAS.2012.8](https://doi.org/10.1109/NAS.2012.8).
- Q. Wang, K. Saha, E. Gregori, D. Joyner, and A. Goel. 2021. "Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant." In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.
- F. R. Ward. 2023. "Defining Deception in Structural Causal Games." en.
- H. de Weerd, R. Verbrugge, and B. Verheij. May 2022. "Higher-order theory of mind is especially useful in unpredictable negotiations." en. *Autonomous Agents and Multi-Agent Systems*, 36, 2, (May 2022), 30. doi:[10.1007/s10458-022-09558-6](https://doi.org/10.1007/s10458-022-09558-6).

- Y. Wen, Y. Yang, R. Luo, J. Wang, and W. Pan. Mar. 2019. *Probabilistic Recursive Reasoning for Multi-Agent Reinforcement Learning*. arXiv:1901.09207 [cs, stat]. (Mar. 2019). doi:[10.48550/arXiv.1901.09207](https://doi.org/10.48550/arXiv.1901.09207).
- S. Westby and C. Riedl. 2022. “Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach.” Publisher: arXiv Version Number: 4. doi:[10.48550/ARXIV.2208.11660](https://doi.org/10.48550/ARXIV.2208.11660).
- J. Wright and K. Leyton-Brown. July 2010. “Beyond Equilibrium: Predicting Human Behavior in Normal-Form Games.” en. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24, 1, (July 2010), 901–907. doi:[10.1609/aaai.v24i1.7644](https://doi.org/10.1609/aaai.v24i1.7644).
- X. Yu, J. Jiang, H. Jiang, and Z. Lu. Aug. 2021. “Model-Based Opponent Modeling.” en. *arXiv:2108.01843 [cs]*, (Aug. 2021). arXiv: 2108.01843. Retrieved Aug. 10, 2021 from <http://arxiv.org/abs/2108.01843>.
- S. Zamir. 2020. *Bayesian games: Games with incomplete information*. Springer.
- S. Zamir. Jan. 1992. “Chapter 5 Repeated games of incomplete information: Zero-sum.” In: *Handbook of Game Theory with Economic Applications*. Vol. 1. Elsevier, (Jan. 1992), 109–154. doi:[10.1016/S1574-0005\(05\)80008-6](https://doi.org/10.1016/S1574-0005(05)80008-6).

5 Acknowledgments

The authors would like to express their gratitude to Ulrike Hahn, Jakob Foerster, Prashant Doshi, Andreas Hula and Yuval Kochman for their valuable feedback.

6 Appendix

6.1 Derivation of the Value Function Gradient w.r.t Policy

Immediate Effect. We compute the gradient of μ 's value function $V^{\pi_{\mu_k}^t}(b_{\mu_k}(\theta_{v_{k-1}}^t))$ in steps. First, we compute the *immediate* effect: changes to μ 's utility function at time t resulting from changes in $\pi_{\mu_k}^t$

$$\frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^t} [u_{\mu}^t(a_{\mu}^t, a_v^t)]]}{\partial \pi_{\mu_k}^t} = \frac{\partial \sum_{a_{\mu}^t} \sum_{a_v^t} u_{\mu}^t(a_{\mu}^t, a_v^t) P(a_{\mu}^t | \pi_{\mu_k}^t) P(a_v^t | \pi_{v_{k-1}}^t)}{\partial \pi_{\mu_k}^t} \quad (35)$$

Since $\pi_{v_{k-1}}^t$ is independent of $\pi_{\mu_k}^t$ - the gradient of the expected immediate utility is:

$$\sum_{a_{\mu}^t} \left(\frac{\partial P(a_{\mu}^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} \sum_{a_v^t} u_{\mu}^t(a_{\mu}^t, a_v^t) P(a_v^t | \pi_{v_{k-1}}^t) \right) \quad (36)$$

Let $\bar{u}(a_{\mu}^t)$ be the expected utility for μ from playing action a_{μ}^t , averaged over $\pi_{v_{k-1}}^t$:

$$\bar{u}(a_{\mu}^t) = E_{\pi_{v_{k-1}}^t} [u_{\mu}^t(a_{\mu}^t, a_v^t)] \quad (37)$$

Plugging Eq. 37 into Eq. 36 yields:

$$\sum_{a_{\mu}^t} \frac{\partial P(a_{\mu}^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} \bar{u}(a_{\mu}^t) \quad (38)$$

Gradual Effect. As changes in μ 's behaviour are propagated through v 's belief and consequently, v 's behaviour, we express μ 's belief with its full form (Eq. 7) - $b_{\mu_k}(\theta_{v_{k-1}}^{t+1}) = p(b_{v_{k-1}}^t | h^{t-1}) \times p(\psi_v | h^{t-1})$. Since ψ_v is independent of μ 's actions, the gradual changes to μ 's value function result from changes to v 's beliefs.

$$\begin{aligned} \gamma \frac{\partial E_{\pi_{\mu_k}^t} [E_{\pi_{v_{k-1}}^t} [V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))]]}{\partial \pi_{\mu_k}^t} &= \gamma E_{\pi_{v_{k-1}}^t} \left[\frac{\partial E_{\pi_{\mu_k}^t} [V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))]}{\partial \pi_{\mu_k}^t} \right] = \\ &\gamma \sum_{a_v^t} P(a_v^t | \pi_{v_{k-1}}^t) \sum_{a_{\mu}^t} \frac{\partial P(a_{\mu}^t | \pi_{\mu_k}^t)}{\partial \pi_{\mu_k}^t} V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1})) + P(a_{\mu}^t | \pi_{\mu_k}^t) \frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t} \end{aligned} \quad (39)$$

The first equation stems from the independence of v 's policy at time t is of μ 's policy at the same time. The 2nd equation is application of the product rule derivation.

We now expand the term $\frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t}$. Recall (from Eq. 14) that:

$$V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1})) = E_{\pi_{\mu_k}^{t+1}}[E_{a_{v_{k-1}}^{t+1} \sim \hat{\pi}_{v_{k-1}}^{t+1}}(\psi_{v_{k-1}}(b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})) [u_{\mu}(a_{\mu}^{t+1}, a_{v_{k-1}}^{t+1}) + \gamma V^{\pi_{\mu_k}^{t+2}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+2}))]]] \quad (40)$$

Since the actions of v at time $t + 1$ are a function of μ 's actions at time t , we compute the gradient of π_v^{t+1} w.r.t π_{μ}^t :

$$\frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t} = \frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{v_{k-1}}^{t+1}} \frac{\partial \pi_{v_{k-1}}^{t+1}}{\partial \pi_{\mu_k}^t} \quad (41)$$

Since the policy of v is a function of its beliefs, we expand the computation to account for belief updates resulting from μ 's actions:

$$\frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{\mu_k}^t} = \frac{\partial V^{\pi_{\mu_k}^{t+1}}(b_{\mu_k}(\theta_{v_{k-1}}^{t+1}))}{\partial \pi_{v_{k-1}}^{t+1}} \frac{\partial \pi_{v_{k-1}}^{t+1}}{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})} \frac{\partial b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1})}{\partial \pi_{\mu_k}^t} \quad (42)$$

In addition, v 's future beliefs are a recurrent function of its current beliefs and μ 's actions: $b_{v_{k-1}}(\theta_{\mu_{k-2}}^{t+1}) = f(b_{v_{k-1}}(\theta_{\mu_{k-2}}^t), a_{\mu}^t)$ (from Eq. 6) - indicating changes to v 's beliefs are propagated into its future beliefs, and these are manifested in its actions which governs μ 's utility. Overall, the changes to μ 's policy affect its present and future utilities in both ways.

6.2 Manipulation and Deception

Manipulation is defined as “the act of changing by artful or unfair means so as to serve one’s purpose”. That is, manipulation requires two component: (a) change another agent’s behaviour and (b) that the change is beneficial to the manipulator. The first condition implies that the agent that is manipulated has some baseline (expected) behaviour in the absent of the manipulation and that the manipulation *causes* the victim to change its behaviour. This is illustrated in in [Jaques et al. \(2019\)](#), where the manipulator is rewarded for the KLD between the expected (manipulation free) and manipulation induced behaviour of the victim. The second condition mandates that this change benefits the manipulator. This is a rather wide definition that encompasses various settings—from cooperative (where the victim may also gain from the manipulation) to competitive environments. In this work we focus on manipulation in competitive setting. This means that the manipulator gains excess utility from the manipulation while the victim receives lower reward, had it not being manipulated.

The above definition is missing a key component to successful manipulation—a manipulable victim. Consider the following example — a trickster learns that they can insert a coin with a rod to a vending machine and once the machine outputs a can of soda it pulls back the coin, tricking the machine to act in its favour. If the vending machine had a verification mechanism to detect such trickery ³ (for example by comparing the number of coins in the machine to the number of cans omitted)—the ruse would fail. We conclude that this third component, detection avoidance, implies that either the manipulator learns quicker than the victim (or that the victim is not adapting at all), or it has the capacity to avoid detection by exploiting limited computational resources of the victim.

In this work we show how deception is a sub-class of manipulation, in which the deceiver manipulates the victim by planting false beliefs in the victim’s “mind”. We show how deception meets all requirements of manipulation—induce changes, personal gain and detection avoidance.

³The authors wish to express their rejection of such mischief.

6.3 N-mechanism Activation Analysis in IUG

The activation of the N-mechanism occurs when one of the component is activated. In this section we provide a detailed analysis of this event, as a function of the mechanism's parameters (δ and ω) and the DoM(1) threshold. This analysis is complementary to the figures depicted in the main body (Fig. 5). The following figure depicts the probability that the N-mechanism will fire (y-axis) as a function of the interaction duration (x-axis) and the aforementioned variables. This figure reinforces the findings from the false positive analysis—the strong typicality component Z^1 is more sensitive than the average reward monitoring Z^2 on early trials—when engaging with the low threshold sender. On the other hand, the opposite is true against the high threshold sender. This may be due to the latter's inability (desire) to make high offers, while the former (low threshold) may offer higher partitions (thus respecting the expected average reward) but is prone to repeat the same offer multiple times—triggering the strong typicality component.

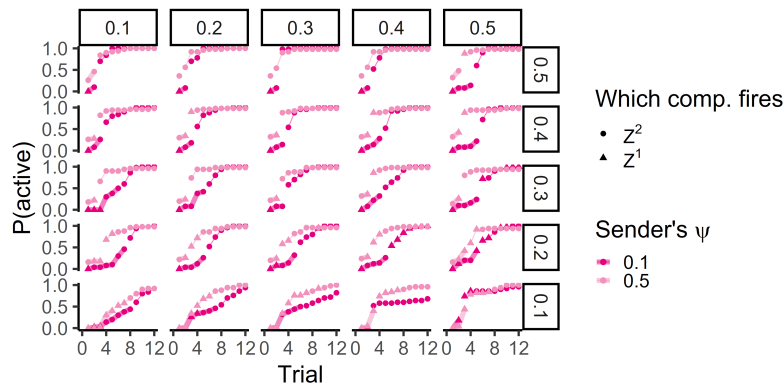


Fig. 9. **Probability of N-mechanism activation by component.** The plot shows the probability that the N-Mechanism is activated as a function of the trial (x-axis), the sender threshold (colour) and the components (point shape). A change in the line shape (thick to thin line) indicates a rise in the N-mechanism triggering probability above 50%. On average, the strong typicality component (Z^1) is more likely to trigger the mechanism early on (marked by triangle point) and later activation of the overall mechanism is typically caused by a deviation from the expected reward (monitored by Z^2)

6.4 Deception in a Zero-Sum Game

Deception is not limited to mixed-motive games, it may also occur in zero-sum games. A canonical example is Poker (Palomäki et al. 2016), in which players deliberately bluff to lure others into increasing the stakes, only to learn in hindsight that they had been tricked. Simple such games were presented and solved by (Zamir 1992). Here, we present an RL variant of one of these games (Example 1.3; Zamir 1992), modelling it with ToM. This illustrates how deceptive behaviour utilises partial information and belief manipulation.

Two agents with different DoM levels play the game presented in Fig. 1 (Row/Column game). In this game, one of two payout matrices G^1, G^2 (Eq. 43) is picked by nature with equal probability, and remains fixed throughout the interaction. The game is played for $T > 0$ trials (12 trials in this work). In each trial the row player picks one of two actions $a_\mu \in \{T, B\}$, corresponding to either the top or bottom row, and the column player picks one of the columns: $a_\nu \in \{L, M, R\}$. The agents pick actions simultaneously and observe the action selected by the opponent before the next trial begins. As in the original paper, the payoffs were hidden and revealed only at the end of the game to avoid disclosing the game played. The entries denote the row player payoff which is generated by the

cell that is selected (the column player gets $r_v = -r_\mu$).

$$G^1 = \begin{pmatrix} 4 & 0 & 2 \\ 4 & 0 & -2 \end{pmatrix}, G^2 = \begin{pmatrix} 0 & 4 & -2 \\ 0 & 4 & 2 \end{pmatrix} \quad (43)$$

The row player (μ) may or may not know which matrix is operational (also with equal probabilities), while the column player (v) is always ignorant of this fact. We denote by $\psi_\mu \in \{0, 1, 2\}$ the row player's persona, with 0 denoting its ignorance of which matrix was picked, and 1, 2 its respective knowledge. Crucially, the agents receive their cumulative reward only at the end of the game rather than throughout. This implies that v can only use μ 's actions to infer this state of the world. Each agent selects its actions using on a Softmax policy (Eq. 3, with a known temperature of 0.1) based on expected discounted long-term reward. In this game we simulate 2 types of row players: either with DoM level $k \in \{-1, 1\}$. We also consider 2 types of v 's DoM level $\{0, 2\}$.

The informed DoM(-1) row player ($\psi_{\mu-1} \in \{1, 2\}$) assumes the column player is uniform $Q_{\mu-1}(a_\mu, \psi_{\mu-1}) = E_{a_v \sim U}[u_\mu(a_\mu, a_v)]$.

The DoM(0) column player infers a posterior belief about the payoff matrix from the actions of the DoM(-1) row player (Eq. 9):

$$b_{v_0}^t(\psi_{\mu-1}) = p(\psi_{\mu-1}|h^{t-1}) \propto P(a_\mu^{t-1}|\psi_{\mu-1})p(\psi_{\mu-1}|h^{t-2}) \quad (44)$$

For example, if the row player constantly plays T , this is a strong signal that the payoff matrix is G^1 , as this action's q-value is $Q_{\mu-1}(T, \psi_{\mu-1} = 1) = 2$ compared to $Q_{\mu-1}(B, \psi_{\mu-1} = 1) = 2/3$. This inference is depicted in Fig. 10(A). Using these beliefs the column player computes the Q-value of each action:

$$Q_{v_0}^*(a_v^t, b_{v_0}^t(\psi_{\mu-1})) = E_{a_\mu \sim \pi_{\mu-1}(\psi_{\mu-1})}[u_v^t(a_\mu^t, a_v^t)] \quad (45)$$

Its policy then favours the column that yields both it and the row player a 0 reward (M in G^1 and L in G^2).

The DoM(1) row player make inferences about the DoM(0) player IRL and plans through its belief update and optimal policy to maximize its reward:

$$Q_{\mu_1}^*(a_\mu^t, \hat{b}_{v_0}^t(\psi_{\mu-1}), \psi_{\mu_1}) = E_{a_v \sim \pi_{v_0}(\theta_{v_0})}[u_v^t(a_\mu^t, a_v^t) + \gamma \max_{a_\mu^{t+1}}\{Q_{\mu_1}^*(a_\mu^{t+1}, \hat{b}_{v_0}^{t+1}(\psi_{\mu-1}), \psi_{\mu_1})\}] \quad (46)$$

notably, these Q-values also take into account the effect of the action on the DoM(0) column player's beliefs.

The DoM(1) policy is to trick the DoM(0) column player into believing a falsehood about the payoff matrix (for example, if the true payoff matrix is G^1 it acts in a way typical for a DoM(-1) in G^2). This deception utilises the same concepts as in the IUG—the limited opponent modelling of the lower DoM column player and its Bayesian IRL (illustrated in Fig. 10(B)). In turn, the DoM(0) column player's q-value computation (Eq. 45) takes as input these false beliefs. This results in its selecting the column that, instead of yielding it a 0 reward, is actually the least favourable column (M in G^1 , L in G^2) yielding it a negative utility of (-4) . This substantially benefits the deceptive DoM(1) row player, as evident in Fig. 10(D) in terms of the difference in reward between row and column players—in this case 7 points on average.

Lastly, the DoM(2) models the row player as a DoM(1). It inverts its actions to make inference about the payoff matrix from and act optimally, similarly to the DoM(0). Using its nested model of the DoM(1) row player, the DoM(2) column player “calls the bluff” and makes correct inferences about the payout matrix Fig. 10(C). Its policy exploits the DoM(1) ruse against itself, by picking the right column, yielding it a reward of 2 and a reward of (-2) to μ_1 . Lacking the capacity to model such counter-deceptive behaviour, the DoM(1) erroneously attributes this behaviour to the SoftMax policy, and its nested beliefs about the column player beliefs are the distorted DoM(0) beliefs. This inability to resist manipulation by the higher DoM column player yields a high income gap, as illustrated in Fig. 10(D).

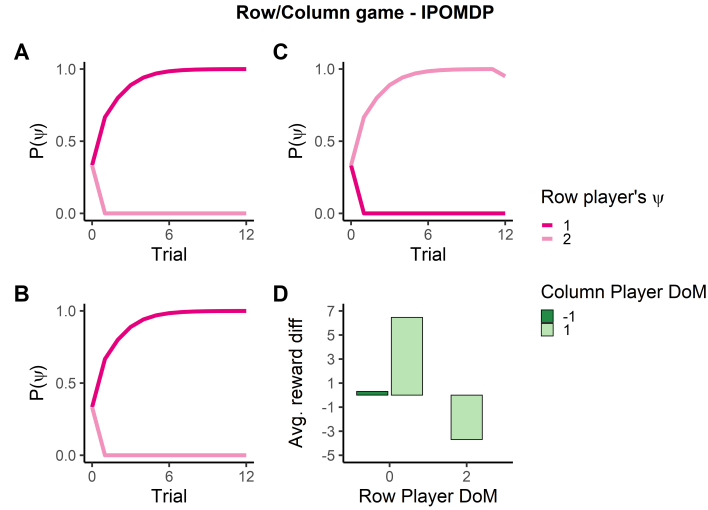


Fig. 10. **Illustration of deception in Row-Column game:** (A) The DoM(0) column player correctly infers the persona $\psi_{\mu-1}$ of the actions of the DoM(-1) row player. In turn its optimal policy yields it a small, positive reward. (B) Taking advantage of its ToM capacity, the DoM(1) row player manipulates the DoM(0) column player's belief by playing a non typical first move but then changing its policy. The deceptive first move causes the DoM(0) column player to form false beliefs and act to the benefit of the deceptive row player. (C) Reading through the bluff of the DoM(1), the DoM(2) column player correctly identifies the DoM(1) row persona and acts in a counter-deceptive manner. (D) The outcome of these manipulations is presented as the average reward difference between the column and row players' rewards.

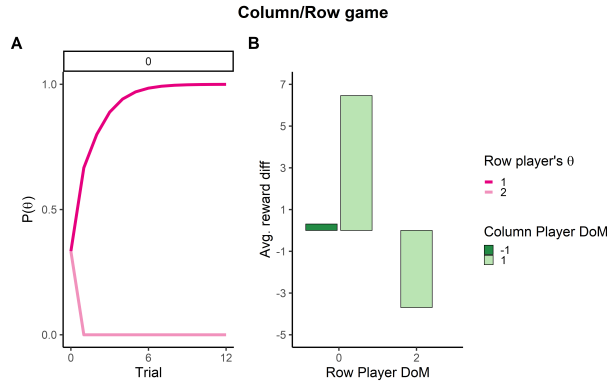


Fig. 11. **Row/Column zero-sum game:** (A) Illustration of DoM(0) Bayesian-IRL. Given the row player's actions, the DoM(0) column player quickly detects the true type (and payoff matrix). (B) Advantage of high DoM level: in each of the simulated dyads, the higher DoM agent has an edge. Its ability to simulate and predict its opponent's behaviour allows it to gain excess wealth

6.4.1 Manipulation, Counter-Manipulation and Counter-Counter Manipulation. The DoM(1) row player's policy exposes it to potential risk, as it plays the "risky" row. For example, in G^1 the right column yields it a negative

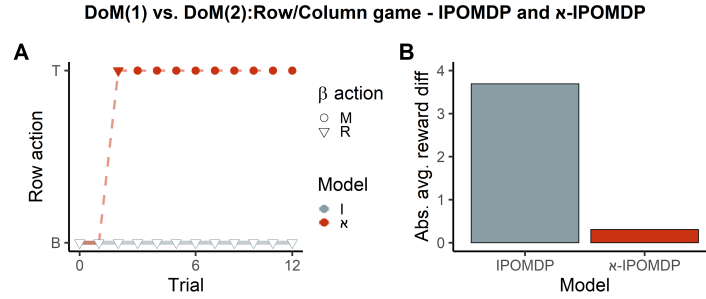


Fig. 12. **Effect of $\aleph$ -IPOMDP in a zero sum game:** (A) Interacting with a $\theta_\mu = 1$, the DoM(0) (left column) is deceived by μ 's actions and form false beliefs. However, the DoM(2) utilises its nested model to read through the bluff and correctly identifies μ 's type. (B)(grey line) In turn, the DoM(2) policy exploits of the DoM(1) ruse against it. Augmented with $\aleph$ -IPOMDP (red line), the DoM(1) infers this unexpected behaviour as a sign of an external entity, triggering the $\aleph$ -mechanism, marked by the dashed line. Its $\aleph$ -policy causes the DoM(2) column player to adapt and alter its abusive behaviour. (C) The effect of the $\aleph$ -IPOMDP in this task is illustrated via the reduction in the average absolute reward difference.

reward. However, given its ability to predict the DoM(0) row player's action this risk is mitigated. As presented above, the DoM(2) column player tricks the trickster by learning to select the right column. The DoM(2) take advantage of the DoM(1) inability to model its behaviour as deceptive and resent it, which yields it a reward of 2 at each trial, depicted in Fig. 11(B). We solve this issue by simulating the game again using the $\aleph$ -IPOMDP framework. Due to the reward masking, the DoM(1) detects that they are matched with an external opponent only through the typical-set component. Due to the expected deterministic behaviour of the column player, we use the δ -typicality algorithm in this task. Identifying that they are outmatched, the $\aleph$ -policy is to play the MinMax policy—selecting the row that yields the highest-lowest reward. Interestingly, in this task, this policy is similar to the optimal policy of the “truth-telling” DoM(−1) agent. In this case the DoM(2) response is to select the column which yields it the highest reward—namely the one that yields it a zero reward, as evident in 12(C).

There is a similar outcome in the zero-sum game. Here, the behaviour of the deceptive DoM(2) column layer described above would be highly non typical for DoM(0) column player, triggering the $\aleph$ -mechanism. The DoM(1) MinMax $\aleph$ -policy, i.e., playing truthfully, causes the DoM(2) to adapt its behaviour appropriately (Fig. 12(B)). In this case, both parties get 0 reward, which drops the average absolute reward difference compared to the IPOMDP case, as illustrated in Fig. 12(C).

Received 20 May 2025; accepted 5 January 2025

Chapter 5

Discussion and Conclusions

This thesis set out to understand the role of Theory of Mind in interactions where deception is beneficial. Across three studies, a coherent and somewhat unsettling picture emerges: ToM is simultaneously the engine of deception, the source of its most dangerous failure modes, and, when properly augmented, a path to its mitigation. I now draw together the threads of the three papers, discuss their limitations, and outline the directions they open.

Summary of Main Findings

The point of departure was the observation that ToM and deception are linked not merely by correlation but by mechanism. *Paper I* [Alon et al., 2023] established this mechanistically: in a Bayesian signaling game, purely reward-maximizing agents endowed with recursive ToM spontaneously discover that distorting their signals is profitable. Crucially, no deceptive intent was hand-coded; the ruse emerged from planning through the observer’s inverse reinforcement learning [Ng et al., 2000]. The information-theoretic analysis provided a precise vocabulary for what “deception” and “skepticism” mean in this context — reduced mutual information between preferences and actions on the sender’s side, and attenuated Bayesian belief updates on the receiver’s side [Kopp et al., 2018]. This result extends and formalizes the Machiavellian ToM hypothesis [Byrne and Whiten, 1988]: the cognitive arms race is not merely a metaphor, it is a provable consequence of reward maximization in games with asymmetric information.

Paper II [Alon et al., 2025] complicated this picture by showing that more ToM is not always better. When an agent’s DoM exceeds what its opponent warrants — the over-mentalizing case — the same Bayesian machinery that correctly identifies a deceptive sender now confabulates deceptive intent from innocuous behavior. The agent develops what amounts to a computational model of paranoia: false beliefs of malicious intent,

rigid updating, and material loss. Critically, the study identified the joint conditions necessary for this failure: behavior must be *ambiguous and perceived as intentional*. Neither condition alone suffices, which is consistent with clinical observations that persecutory delusions involve both hypervigilance to ambiguous social signals *and* the attribution of agency to their source [Freeman, 2016a]. The paper thus repositions paranoia not as a deficit of ToM but as its maladaptive over-application — a distinction with meaningful therapeutic implications.

Paper III [Alon et al., 2026] closed the loop by asking whether there is a path out of the dilemma posed by the first two papers: how can agents with a fixed $\text{DoM}(k)$ reasoning, prone to manipulation by those with a higher $\text{DoM}(k+1)$ level, resist deception, if they lack the means to understand it? The $\aleph$ -IPOMDP answers by separating *detection* from *modeling*: an agent need not understand *how* it is being deceived in order to detect *that* it is being deceived. By monitoring the typicality of observed behavior and the deviation of realized reward from off-policy counterfactual expectations, the $\aleph$ -mechanism flags unmodeled opponents without ascending the cognitive hierarchy. The credible threat encoded in the out-of-belief (OOB) policy then deters the deceiver from executing its full strategy. Notably, an over-sensitive $\aleph$ -mechanism recapitulates the paranoid phenotype of *Paper II*, this time arising from anomaly detection rather than recursive misattribution — two mechanistically distinct routes to the same clinical presentation.

The Calibration Principle — A Path Toward *Adapted* ToM

The unifying thread across all three papers is *calibration*: the alignment between an agent’s DoM and its opponent’s actual level of sophistication (Fig. 5.1). Here, ToM is a precision instrument, enabling both the exploitation of less sophisticated others (deception) and the resistance to more sophisticated ones (skepticism). The red cells are the failure zone: over-mentalizing produces paranoia; under-mentalizing produces victimhood.

This framing implies that the question “how much ToM is enough?” has no universal answer. The optimal level is *opponent-relative*. In environments populated by agents of varying sophistication — as human social environments invariably are [Camerer et al., 2004, Nagel, 1995] — an agent benefits from the ability to estimate its opponent’s DoM and adapt accordingly [Yu et al., 2022, Alaoui and Penta, 2016]. *Paper III* provides a partial solution: the $\aleph$ -mechanism detects when observed others do not as predicted by the current DoM. But this solution is insufficient without requiring the agent to estimate the opponent’s true level. Future work may combine this with online DoM estimation to produce a fully adaptive, intentional, system [Roth and Marsella, 2026,

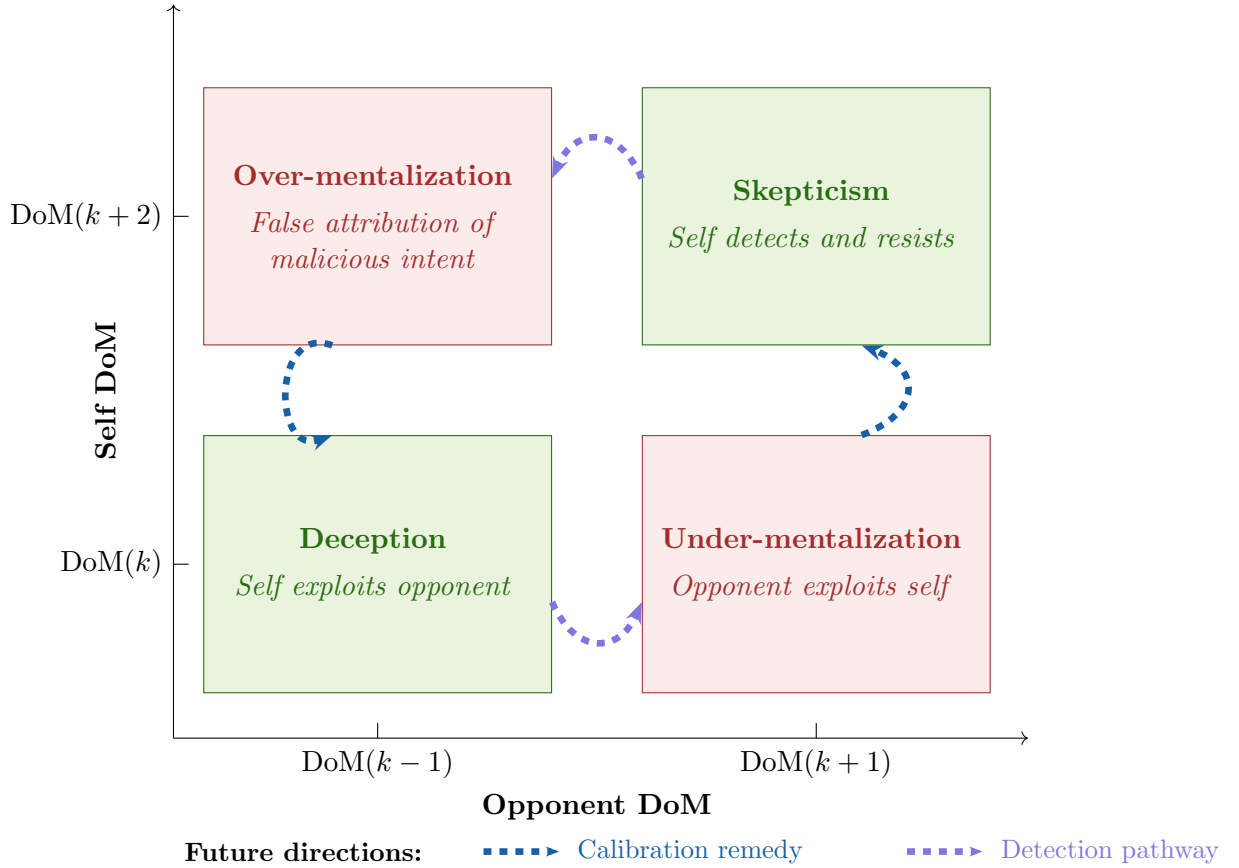


Figure 5.1: **From ToM Miscalibration to Adapted ToM:** Green cells (main diagonal) show calibrated outcomes; red cells (off-diagonal) show failure modes. Blue arrows indicate *adaptation* remedies: over-mentalizing agents that learn to identify and reduce their DoM avoid over-mentalization, and under-mentalizing agents augment their DoM level to avoid exploitation. Purple arrows mark *miscalibration identification* — agents infer that their DoM level is miscalibrated relative to the opponent.

Wagner et al., 2025].

Limitations

Several limitations qualify the conclusions of this thesis.

Strict K-level assumption. All three papers adopt the strict K-level model, in which a $\text{DoM}(k)$ agent models all opponents as $\text{DoM}(k-1)$. This is a known idealization [Camerer et al., 2004]. The Cognitive Hierarchy model [Camerer et al., 2004], in which agents hold a distribution over opponent DoM levels, would allow for more graceful degradation when the true opponent level is uncertain. Extending the $\aleph$ -IPOMDP to a mixture-model opponent representation [Camerer et al., 2004] is a natural next step; it would also reduce the hard binary trigger of the $\aleph$ -mechanism, replacing it with a probabilistic “goodness of fit” measure more consistent with human uncertainty

reasoning [Oey et al., 2023].

Simplified utility structures. *Paper I* uses a T-Maze paradigm with additive utilities; *Paper II* and *Paper III* use threshold utilities in the Iterated Ultimatum Game (IUG). Both abstractions remove social preferences — inequity aversion [Fehr and Schmidt, 1999, Hula et al., 2018], spite, altruism — that are empirically important in human bargaining. Future work should embed the deception framework within richer utility structures, such as Fehr-Schmidt utilities, to test whether the calibration principle holds when social preferences create additional incentives for or against deception.

Simulated agents, not human participants. The three papers are exclusively computational. The models generate testable predictions — particularly *Paper II*’s prediction that high trait paranoia correlates with mismatched DoM in the IUG — but these remain unverified against human data. Empirical validation using computational psychiatry paradigms [Buergi et al., 2024, Hula et al., 2018], ideally with clinical populations, is the most pressing empirical priority.

No multi-agent learning. All agents in this thesis use fixed policies computed from prior knowledge of the game structure and adhere to a closed-set representation (all types are known and inferred). In reality, agents learn over repeated interactions, and deceptive strategies may evolve or erode as a function of experience [Sarkadi, 2023]. Integrating the $\aleph$ -mechanism with online learning — so that the typical-set model is updated from experience rather than fixed at initialization — would make the framework more realistic. In turn, this will likely uncover other deceptive dynamics, like opponent shaping [Foerster et al., 2018, Jaques et al., 2019].

Future Directions

From computation to clinic The most direct extension of *Paper II* is a human experiment using the IUG with computerized senders spanning DoM levels from $k = -1$ to $k = 2$. Participants with high trait paranoia, and patients with persecutory delusions or borderline personality disorder, should be more likely to exhibit mismatched DoM against simple senders — applying DoM(1) or DoM(2) reasoning to a DoM(−1) opponent — and should be slower to correct this mismatch even when evidence accumulates. The D2/D3 dopamine system has been implicated in the precision of social inference [Barnby et al., 2024]; pharmacological manipulation could dissociate the over-mentalizing route to paranoia from the $\aleph$ -sensitivity route identified in *Paper III*.

ToM in large language models The emergence of ToM-like capacities in large language models (LLMs) [Kosinski, 2023] and the concomitant risk of deceptive behavior [Hagendorff, 2024] make the frameworks developed here directly relevant to AI safety. The $\aleph$ -IPOMDP offers a blueprint for an external monitoring layer: a system that compares an LLM’s outputs against the outputs expected of a well-aligned model, flags systematic deviations, and triggers a defensive policy (e.g. routing to a human reviewer). Formalizing LLM interaction as a Bayesian repeated game and applying the $\aleph$ -mechanism to it is a concrete near-term research program.

Adaptive DoM and Metacognition A natural extension of all three papers is to make DoM an adaptive parameter rather than a fixed one. Agents that can detect when their current DoM is miscalibrated and revise it accordingly — essentially a metacognitive layer over the K-level model [Schulz et al., 2023a] — would avoid both the exploitation risk of under-mentalizing and the paranoid failure mode of over-mentalizing. Such a system would need a stopping criterion: when is the evidence for a DoM shift strong enough to warrant the computational cost? The $\aleph$ -mechanism provides a natural trigger for upward revision (anomaly detection); a symmetric downward revision mechanism, triggered by consistently over-predicted opponent sophistication, would complete the picture.

Closing Remarks

This thesis began with a simple observation: we learn to deceive, and then we learn that others can deceive us. The three papers trace the formal consequences of this observation through the lens of recursive opponent modeling. Deception is not a special faculty; it is what reward-maximizing agents with ToM *do*. Paranoia is not a deficit; it is what those same agents suffer when their ToM is pointed at the wrong target. And the path out is not more recursion but a principled acknowledgment of its limits: knowing that one does not know, and acting accordingly.

Bibliography

- Matthew Aitchison, Lyndon Benke, and Penny Sweetser. Learning to Deceive in Multi-agent Hidden Role Games. In Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney, editors, *Deceptive AI*, Communications in Computer and Information Science, pages 55–75, Cham, 2021. Springer International Publishing. ISBN 978-3-030-91779-1. doi: 10.1007/978-3-030-91779-1_5.
- Larbi Alaoui and Antonio Penta. Endogenous Depth of Reasoning. *The Review of Economic Studies*, 83(4):1297–1333, October 2016. ISSN 0034-6527. doi: 10.1093/restud/rdv052. URL <https://doi.org/10.1093/restud/rdv052>.
- Stefano V. Albrecht and Peter Stone. Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. *Artificial Intelligence*, 258:66–95, May 2018. ISSN 00043702. doi: 10.1016/j.artint.2018.01.002. URL <http://arxiv.org/abs/1709.08071>. arXiv: 1709.08071.
- Nitay Alon, Lion Schulz, Jeffrey S. Rosenschein, and Peter Dayan. A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind*, 7:608–624, August 2023. ISSN 2470-2986. doi: 10.1162/opmi_a.00097. URL https://doi.org/10.1162/opmi_a_00097.
- Nitay Alon, Lion Schulz, Vaughan Bell, Michael Moutoussis, Peter Dayan, and Joseph M. Barnby. (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1):159–177, September 2025. ISSN 2379-6227. doi: 10.5334/cpsy.117. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11396085/>.
- Nitay Alon, Joseph M. Barnby, Stefan Sarkadi, Lion Schulz, Jeffrey S. Rosenschein, and Peter Dayan. 8-IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection. *Journal of Artificial Intelligence Research*, 85, February 2026. ISSN 1076-9757. doi: 10.1613/jair.1.19204. URL <https://www.jair.org/index.php/jair/article/view/19204>.

- Ian A Apperly. What is “theory of mind”? concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology*, 65(5):825–839, 2012.
- Ayala Arad and Ariel Rubinstein. The 11-20 money request game: Evaluating the upper bound of k-level reasoning. *Eitan Fiberglas School of Economics, Tel Aviv*, 2010.
- Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33), 2011. URL <https://escholarship.org/uc/item/5rk7z59q>.
- Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, December 2009. ISSN 00100277. doi: 10.1016/j.cognition.2009.07.005. URL <https://linkinghub.elsevier.com/retrieve/pii/S0010027709001607>.
- Mahzarin R Banaji and Susan A Gelman. *Navigating the social world: What infants, children, and other species can teach us*. OUP USA, 2013.
- Deanna M. Barch and Andrew Ceaser. Cognition in schizophrenia: Core psychological and neural mechanisms. *Trends in Cognitive Sciences*, 16(1):27–34, 2012. doi: 10.1016/j.tics.2011.11.015.
- JM Barnby, V Bell, Q Deeley, MA Mehta, and M Moutoussis. D2/d3 dopamine supports the precision of mental state inferences and self-relevance of joint social outcomes. *Nature Mental Health*, 2(5):562–573, 2024.
- Joseph M Barnby, Peter Dayan, and Vaughan Bell. Formalising social representation to explain psychiatric symptoms. *Trends in Cognitive Sciences*, 2023.
- Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a “theory of mind” ? *Cognition*, 21(1):37–46, October 1985. ISSN 0010-0277. doi: 10.1016/0010-0277(85)90022-8. URL <https://www.sciencedirect.com/science/article/pii/0010027785900228>.
- Anthony Bateman and Peter Fonagy. *Mentalization-Based Treatment for Borderline Personality Disorder: A Practical Guide*. Oxford University Press, 2010.
- Volker Benndorf, Dorothea Kübler, and Hans-Theo Normann. Depth of reasoning and information revelation: An experiment on the distribution of k-levels. *International Game Theory Review*, 19(04):1750021, 2017.
- Peter G. Bennett. Hypergames: Developing a model of conflict. *Futures*, 12(6):489–507, December 1980. ISSN 0016-3287. doi: 10.1016/0016-3287(80)90005-1. URL <https://www.sciencedirect.com/science/article/pii/0016328780900051>.

- Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research*, 27(4):819–840, November 2002. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.27.4.819.297. URL <https://pubsonline.informs.org/doi/10.1287/moor.27.4.819.297>.
- Rahul Bhui, Lawrence Lai, and Samuel J. Gershman. Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41:15–21, 2021. doi: 10.1016/j.cobeha.2021.02.015.
- Lois Bloom. *The transition from infancy to language: Acquiring the power of expression*. Cambridge University Press, 1993.
- Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, January 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao1733. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aao1733>.
- Niklas Buergi, Gökhan Aydogan, Arkady Konovalov, and Christian Ruff. A neural fingerprint of adaptive mentalization. February 2024. doi: 10.31234/osf.io/npqy8. URL <https://osf.io/npqy8>.
- Richard W. Byrne and Andrew Whiten, editors. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press, Oxford, New York, July 1988. ISBN 978-0-19-852175-4.
- Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A Cognitive Hierarchy Model of Games*. *The Quarterly Journal of Economics*, 119(3):861–898, August 2004. ISSN 0033-5533. doi: 10.1162/0033553041502225. URL <https://doi.org/10.1162/0033553041502225>.
- Stephanie M. Carlson and Louis J. Moses. Individual differences in inhibitory control and children’s theory of mind. *Child Development*, 72(4):1032–1053, 2001. doi: 10.1111/1467-8624.00333.
- Michael Chandler, Anna S Fritz, and Suzanne Hala. Small-scale deceit: Deception as a marker of two-, three-, and four-year-olds’ early theories of mind. *Child development*, pages 1263–1277, 1989.
- Roderick M. Chisholm and Thomas D. Feehan. The intent to deceive. *The Journal of Philosophy*, 74(3):143–159, 1977. doi: 10.2307/2025605.
- Vincent P Crawford. Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 93(1):133–149, 2003.

- Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31(2):250–287, March 2017. ISSN 1573-7454. doi: 10.1007/s10458-015-9317-1. URL <https://doi.org/10.1007/s10458-015-9317-1>.
- Rory T. Devine and Claire Hughes. Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, 87(6): 1777–1794, 2016. doi: 10.1111/cdev.12589.
- Xiao Pan Ding, Henry M. Wellman, Yu Wang, Genyue Fu, and Kang Lee. Theory-of-Mind Training Causes Honest Young Children to Lie. *Psychological Science*, 26(11): 1812–1821, November 2015. ISSN 0956-7976. doi: 10.1177/0956797615604628. URL <https://doi.org/10.1177/0956797615604628>.
- Prashant Doshi and Piotr J Gmytrasiewicz. Monte carlo sampling methods for approximating interactive pomdps. *Journal of Artificial Intelligence Research*, 34:297–337, 2009.
- David Ettinger and Philippe Jehiel. A Theory of Deception. *American Economic Journal: Microeconomics*, 2(1):1–20, February 2010. ISSN 1945-7669. doi: 10.1257/mic.2.1.1. URL <https://www.aeaweb.org/articles?id=10.1257/mic.2.1.1>.
- Ernst Fehr and Klaus M. Schmidt. A Theory of Fairness, Competition, and Cooperation*. *The Quarterly Journal of Economics*, 114(3):817–868, August 1999. ISSN 0033-5533. doi: 10.1162/003355399556151. URL <https://doi.org/10.1162/003355399556151>.
- Stephen M. Fleming and Raymond J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B*, 367(1594):1338–1349, 2012. doi: 10.1098/rstb.2011.0417.
- Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with Opponent-Learning Awareness. *arXiv:1709.04326 [cs]*, September 2018. URL <http://arxiv.org/abs/1709.04326>. arXiv: 1709.04326.
- Peter Fonagy and Anthony W. Bateman. Mechanisms of change in mentalization-based treatment of BPD. *Journal of Clinical Psychology*, 62(4):411–430, April 2006. ISSN 0021-9762. doi: 10.1002/jclp.20241.
- M. C. Frank and N. D. Goodman. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998, May 2012. ISSN 0036-8075, 1095-9203. doi: 10.

- 1126/science.1218633. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1218633>.
- Daniel Freeman. Persecutory delusions: a cognitive perspective on understanding and treatment. *The lancet psychiatry*, 3(7):685–692, 2016a.
- Daniel Freeman. Persecutory delusions: A cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7):685–692, 2016b. doi: 10.1016/S2215-0366(16)00066-3.
- Daniel Freeman, Philippa A. Garety, Paul E. Bebbington, Benjamin Smith, Rebecca Rollinson, David Fowler, Elizabeth Kuipers, Katarzyna Ray, and Graham Dunn. Psychological investigation of the structure of paranoia in a non-clinical population. *British Journal of Psychiatry*, 186(5):427–435, 2005. doi: 10.1192/bjp.186.5.427.
- Chris D. Frith. The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B*, 367(1599):2213–2223, 2012. doi: 10.1098/rstb.2012.0123.
- Uta Frith and Francesca Happé. Autism: Beyond “theory of mind”. *Cognition*, 50(1–3): 115–132, 1994. doi: 10.1016/0010-0277(94)90024-8.
- Philip S Gallo Jr and Charles G McClintock. Cooperative and competitive behavior in mixed-motive games. *Journal of Conflict Resolution*, 9(1):68–78, 1965.
- Philippa A. Garety, Elizabeth Kuipers, David Fowler, Daniel Freeman, and Paul E. Bebbington. A cognitive model of the positive symptoms of psychosis. *Psychological Medicine*, 31(2):189–195, 2001. doi: 10.1017/S0033291701003312.
- Vanessa E. Ghosh and Asaf Gilboa. What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53:104–114, 2014. doi: 10.1016/j.neuropsychologia.2013.11.010.
- P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24:49–79, July 2005. ISSN 1076-9757. doi: 10.1613/jair.1579. URL <https://jair.org/index.php/jair/article/view/10414>.
- Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International conference on autonomous agents and multiagent systems*, pages 66–83. Springer, 2017.
- Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

- John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part iii. the basic probability distribution of the game. *Management Science*, 14(7):486–502, 1968. ISSN 00251909, 15265501. URL <http://www.jstor.org/stable/2628894>.
- Mark K. Ho, Rebecca Saxe, and Fiery Cushman. Planning with Theory of Mind. *Trends in Cognitive Sciences*, September 2022. ISSN 1364-6613. doi: 10.1016/j.tics.2022.08.003. URL <https://www.sciencedirect.com/science/article/pii/S1364661322001851>.
- Teck-Hua Ho and Xuanming Su. A dynamic level-k model in sequential games. *Management Science*, 59(2):452–469, 2013.
- Trong Nghia Hoang and Kian Hsiang Low. Interactive POMDP Lite: Towards Practical Planning to Predict and Exploit Intentions for Interacting with Self-Interested Agents. In *Twenty-Third International Joint Conference on Artificial Intelligence*, June 2013. URL <https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6207>.
- Paktwis Hodayun. Negotiating with Deceptive Agents in Mixed-Motive Interaction Environments using Theory of Mind. page 99, August 2018.
- Linan Huang and Quanyan Zhu. Dynamic bayesian games for adversarial and defensive cyber deception. *ArXiv*, abs/1809.02013, 2018. URL <https://api.semanticscholar.org/CorpusID:52171834>.
- Andreas Hula, P. Read Montague, and Peter Dayan. Monte Carlo Planning Method Estimates Planning Horizons during Interactive Social Exchange. *PLOS Computational Biology*, 11(6):e1004254, June 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004254. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004254>.
- Andreas Hula, Iris Vilares, Terry Lohrenz, Peter Dayan, and P. Read Montague. A model of risk and mental state shifts during social interaction. *PLOS Computational Biology*, 14(2):e1005935, February 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005935. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005935>.
- Quentin J. M. Huys, Tiago V. Maia, and Michael J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016. doi: 10.1038/nn.4238.
- Arnold Isenberg. Conditions for lying. *Ethical Theory and Business*, 1973. Reprinted in *Aesthetics and the Theory of Criticism*, University of Chicago Press, pp. 248–253.

- Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. *arXiv:1810.08647 [cs, stat]*, June 2019. URL <http://arxiv.org/abs/1810.08647>. arXiv: 1810.08647.
- Julian Jara-Ettinger, Hyowon Gweon, Laura E. Schulz, and Joshua B. Tenenbaum. The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, 20(8):589–604, August 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.05.011. URL <https://linkinghub.elsevier.com/retrieve/pii/S1364661316300535>.
- Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1): 99–134, May 1998. ISSN 0004-3702. doi: 10.1016/S0004-3702(98)00023-X. URL <http://www.sciencedirect.com/science/article/pii/S000437029800023X>.
- Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. Hypothesis-Driven Theory-of-Mind Reasoning for Large Language Models, February 2025. URL <http://arxiv.org/abs/2502.11881>. arXiv:2502.11881 [cs].
- Brooks King-Casas, Carla Sharp, Laura Lomax-Bream, Terry Lohrenz, Peter Fonagy, and P. Read Montague. The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890):806–810, 2008. doi: 10.1126/science.1156902.
- Max Kleiman-Weiner, Alejandro Vientós, David G. Rand, and Joshua B. Tenenbaum. Evolving general cooperation with a Bayesian theory of mind. *Proceedings of the National Academy of Sciences*, 122(25):e2400993122, June 2025. doi: 10.1073/pnas.2400993122. URL <https://www.pnas.org/doi/full/10.1073/pnas.2400993122>.
- C. Kopp. Shannon, hypergames and information warfare. In W. Hutchinson, editor, *Proceedings of the 3rd Australian Information Warfare & Security Conference (IWAR ’02)*, Perth, Australia, November 2002.
- Carlo Kopp, Kevin B. Korb, and Bruce I. Mills. Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to “fake news”. *PLOS ONE*, 13(11):e0207383, November 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207383. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207383>.
- Michal Kosinski. Theory of Mind May Have Spontaneously Emerged in Large Language Models, February 2023. URL <http://arxiv.org/abs/2302.02083>. arXiv:2302.02083 [cs].

- Nicholas S. Kovach, Alan S. Gibson, and Gary B. Lamont. Hypergame Theory: A Model for Conflict, Misperception, and Deception. *Game Theory*, 2015:e570639, August 2015. ISSN 2356-6930. doi: 10.1155/2015/570639. URL <https://www.hindawi.com/journals/gt/2015/570639/>.
- Paul H. Lysaker and Giancarlo Dimaggio. Metacognitive capacities for reflection in schizophrenia: Implications for developing treatments. *Schizophrenia Bulletin*, 40(3): 487–491, 2014. doi: 10.1093/schbul/sbu038.
- David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- James Edwin Mahon. The definition of lying and deception. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018. URL <https://plato.stanford.edu/archives/fall2018/entries/lying-definition/>.
- Michael Maschler, Shmuel Zamir, and Eilon Solan. *Game theory*. Cambridge University Press, 2020.
- Peta Masters and Sebastian Sardina. Deceptive path-planning. In *IJCAI*, pages 4368–4375, 2017.
- Andrew N Meltzoff. Origins of theory of mind, cognition and communication. *Journal of communication disorders*, 32(4):251–269, 1999.
- Peter Mitchell. *Acquiring a conception of mind: A review of psychological research and theory*. Psychology Press, 2021.
- Shuwa Miura and Shlomo Zilberstein. Observer-Aware Planning with Implicit and Explicit Communication. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS '24*, pages 1409–1417, Richland, SC, May 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 979-8-4007-0486-4.
- Mikael Mikkonen and Carita Lindstedt. The evolutionary ecology of deception. *Biological Reviews*, 91(4):1020–1035, 2016.
- Henrike Moll and Michael Tomasello. Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):639–648, April 2007. ISSN 0962-8436. doi: 10.1098/rstb.2006.2000. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2346522/>.

- P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012. doi: 10.1016/j.tics.2011.11.018.
- Rosemarie Nagel. Unraveling in Guessing Games: An Experimental Study. *The American Economic Review*, 85(5):1313–1326, 1995. ISSN 0002-8282. URL <https://www.jstor.org/stable/2950991>.
- Thomas O. Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26:125–173, 1990. doi: 10.1016/S0079-7421(08)60053-5.
- Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.
- Mike Oaksford and Nick Chater. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007.
- Lauren A Oey, Adena Schachner, and Edward Vul. Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 152(2):346, 2023.
- Eric Pacuit and Olivier Roy. Epistemic Foundations of Game Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017. URL <https://plato.stanford.edu/archives/sum2017/entries/epistemic-game/>.
- David L. Penn, Lawrence J. Sanna, and David L. Roberts. Social cognition in schizophrenia: An overview. *Schizophrenia Bulletin*, 34(3):408–411, 2008. doi: 10.1093/schbul/sbn014.
- Josef Perner and Birgit Lang. Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3(9):337–344, 1999. doi: 10.1016/S1364-6613(99)01362-5.
- Stefan Poslad. Specifying protocols for multi-agent systems interaction. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 2(4):15–es, 2007.
- David Premack and Guy Woodruff. 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/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0>.

- Miquel Ramirez and Hector Geffner. Plan recognition as planning. pages 1778–1783, 2009.
- Debajyoti Ray, Brooks King-Casas, P Montague, and Peter Dayan. Bayesian model of behaviour in economic games. *Advances in neural information processing systems*, 21, 2008.
- Marion Roth and Stacy Marsella. Theory of mind on demand: do we prepare or react? *Frontiers in Psychology*, 17, February 2026. ISSN 1664-1078. doi: 10.3389/fpsyg.2026.1693027. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2026.1693027/full>.
- Stefan Sarkadi. An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies. In *4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems ACSOS 2023*. IEEE Computer Society, 2023.
- Hampei Sasahara and Henrik Sandberg. Epistemic Signaling Games for Cyber Deception With Asymmetric Recognition. *IEEE Control Systems Letters*, 6:854–859, 2022. ISSN 2475-1456. doi: 10.1109/LCSYS.2021.3087097. URL <https://ieeexplore.ieee.org/document/9447822/>.
- Jennifer Saul. *Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford University Press, 2012.
- Lion Schulz, Stephen M Fleming, and Peter Dayan. Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3):604, 2023a.
- Lion Schulz, Stephen M. Fleming, and Peter Dayan. Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3):604–639, 2023b. doi: 10.1037/rev0000401.
- Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, June 2014. ISSN 00100285. doi: 10.1016/j.cogpsych.2013.12.004. URL <https://linkinghub.elsevier.com/retrieve/pii/S0010028514000024>.
- Carla Sharp, Hilary Pane, Christina Ha, Amanda Venta, Aakash B. Patel, Jennifer Sturek, and Peter Fonagy. 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. doi: 10.1016/j.jaac.2011.01.017.
- David Silver and Joel Veness. Monte-Carlo Planning in Large POMDPs. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances*

- in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/edfbelafcf9246bb0d40eb4d8027d90f-Paper.pdf.
- Kamila E. Sip, Andreas Roepstorff, William McGregor, and Chris D. Frith. Detecting deception: the scope and limits. *Trends in Cognitive Sciences*, 12(2):48–53, February 2008. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2007.11.008. URL [https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613\(07\)00335-X](https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(07)00335-X).
- Joel Sobel. Signaling games. In *Complex social and behavioral systems: Game theory and agent-based models*, pages 251–268. Springer, 2020.
- Beate Sodian, Catherine Taylor, Paul L Harris, and Josef Perner. Early deception and the child’s theory of mind: False trails and genuine markers. *Child development*, 62(3):468–483, 1991.
- Finnegan Southey, Michael P. Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes’ Bluff: Opponent Modelling in Poker, July 2012. URL <http://arxiv.org/abs/1207.1411>. arXiv:1207.1411 [cs].
- Michael Spence. Job market signaling. In *Uncertainty in economics*, pages 281–306. Elsevier, 1978.
- Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1504–1509, July 2010. ISSN 2374-3468. doi: 10.1609/aaai.v24i1.7529. URL <https://ojs.aaai.org/index.php/AAAI/article/view/7529>.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.
- John Tsitsiklis and Michael Athans. On the complexity of decentralized decision making and detection problems. *IEEE Transactions on Automatic Control*, 30(5):440–446, 2003.
- Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-Set Recognition: a Good Closed-Set Classifier is All You Need? April 2022. doi: 10.48550/arXiv.2110.06207. URL <http://arxiv.org/abs/2110.06207>. arXiv:2110.06207 [cs].
- Lev Semenovich Vygotsky and Michael Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

- Eitan Wagner, Nitay Alon, Joseph M. Barnby, and Omri Abend. Mind Your Theory: Theory of Mind Goes Deeper Than Reasoning, June 2025. URL <http://arxiv.org/abs/2412.13631>. arXiv:2412.13631 [cs].
- Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. Defining deception in structural causal games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2902–2904, 2023.
- Michael Wooldridge and Nicholas R Jennings. Agent theories, architectures, and languages: a survey. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–39. Springer, 1994.
- Ting Xiang, Terry Lohrenz, and P Read Montague. Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, 33(3):1099–1108, 2013.
- Wako Yoshida, Ray J. Dolan, and Karl J. Friston. Game Theory of Mind. *PLOS Computational Biology*, 4(12):e1000254, December 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000254. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000254>.
- Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. *Advances in Neural Information Processing Systems*, 35: 28208–28221, 2022.
- Shmuel Zamir. Bayesian games: Games with incomplete information. In *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 119–137. Springer, 2020.

שברשותו, ואם התגמולים שהוא מקבל תואמים לתחזיות של אותם מודלים. כאשר התשובה לאחת מהשאלות היא שלילית, הקורבן יכול להסיק שהוא מתמודד עם יריב לא-ידוע ולעבור לעמדת הגנה שנועדה להפוך את ההונאה מצעד שמקנה יתרון למהלך שמוביל לכאב הדדי. אנו מראים כי תוספת פשוטה זו מצמצמת באופן משמעותי את היתרון של מטעה מתוחכם על קורבן מוגבל, מבלי לדרוש מן הקורבן לרמת הסקה גבוהה יותר.

שלוש התרומות יחדיו מתכנסות לעיקרון מארגן יחיד: המשתנה המכריע הקובע אם תיאורייה של התודעה היא אדפטיבית או פתולוגית אינו העומק המוחלט של עומק ההסקה הרקורסיבית, אלא הכיול שלו ביחס ליריב. כאשר הכיול מדויק, תיאורייה של התודעה היא כלי מדויק המאפשר הטעייה אסטרטגית וספקנות מוצדקת במידה שווה. כאשר היא חורגת יתר על המידה, היא הופכת למכונת פרנויה. כאשר היא נמוכה מדי, היא משאירה את הסוכן חשוף לניצול. המסגרת החישובית שפותחה כאן — ובפרט התובנה כי סוכנים יכולים לדעת שאינם יודעים, ולנהוג בהתאם — טומנת בחובה השלכות על הבנתנו את ההפרעות הנפשיות הנובעות מייחוס חברתי שגוי, על עיצוב מערכות בינה מלאכותית המתקשרות עם בני אדם אסטרטגיים, ועל השאלה הרחבה יותר כיצד סוכנים חכמים יכולים להישאר חסינים בעולם המאוכלס באחרים המנסים לתמרן אותם.

תקציר

רמאות היא תופעה קבועה בחיים חברתיים. מרוכש דירה המתחזה לאדיש כדי להוריד את המחיר המבוקש, ועד לשחקן פוקר המעמיד פנים שיש לו יד חלשה כדי לפתות שחקנים אחרים להגדיל את סכום ההימור — סוכנים מעוותים באופן שגרתי את האותות שהם משדרים ומתעלמים בצורה אסטרטגית מהאותות שהם קולטים. הונאה אינה כישרון מיוחד; היא פשוט מה שסוכנים בעלי תחכום אסטרטגי עושים כאשר האינטרסים שלהם מתנגשים והמידע אינו סימטרי. עבודת מחקר זו בוחנת את המנגנון הקוגניטיבי והחישובי המאפשר רמאות, שואלת מה קורה כשמנגנון זה נכשל, ומציעה דרך עקרונית לאיתור הונאה ולהרתעה ממנה — מבלי להיכנס למירוץ חימוש של קוגניטיב.

המרכיב הקוגניטיבי המרכזי הוא תיאוריה של התודעה (Theory of Mind) — היכולת לייצג את האמונות, הרצונות והכוונות של אחרים, כולל האמונות שלהם לגבי תודעה של האחד עצמו. כאשר סוכן יכול לצפות כיצד צופה יפרש את פעולותיו, הוא מקבל את הכוח לעצב פעולות שמטמיעות אמונות שגויות אצל אותו צופה. יכולת זו היא רקורסיבית: תודעה מתוחכמת דייה יכול למדל יריב שהוא עצמו ממדל את בעל התודעה, וכן הלאה — דרך שכבות של "אני חושב שאתה חושב שאני חושב". עבודה זו מראה כי רקורסיה זו היא גם מנוע ההטעייה וגם מקור כשלי הדרך המסוכנים ביותר שלה.

התרומה הראשונה מבססת כי הטעייה וספקנות צצות מתוך מיקסום תועלת, כאשר סוכנים ניחנו בתיאוריה (רקורסיבית) של התודעה, ללא כל תכנות מכון של הונאה. במשחק איתות שבו בחירות הקונה חושפות מידע שמוכר ישתמש בו לקביעת מחירים, קונה בעל תחכום מספיק, מגלה כי עיוות בחירותיו הוא כדאי. התחבולה אינה מתוכננת; היא צומחת כתוצאה טבעית מתכנון דרך תהליך ההסקה של הצופה. מנגד, מוכר מתוחכם יותר, שמודע לאפשרות זו, לומד להפחית ממשקל האותות שהקונה שולח — ונהיה ספקן באופן אסטרטגי. תוך שימוש בתורת המידע למדידת דינמיקות אלו, אנו מראים בדיוק כמה מידע הסוכן המטעה מסתיר וכמה הסוכן הספקני מפחית ממשקלו, ומספקים גשר פורמלי בין התנהגות אסטרטגית לבין האמונות שהיא נועדה לתמרן.

התרומה השנייה מסבכת את המסקנה הטבעית הנובעת מן הראשונה. אם הסקה רקורסיבית עמוקה יותר מאפשרת הטעייה ואיתורה, האם על סוכנים פשוט למקסם את עומק ההיגיון שלהם? התשובה, כפי שמתברר, היא שלילית. כאשר סוכן מיישם מודל מתוחכם יותר ממה שיריבו מצדיק בפועל, אותו מנגנון הסקה שבמצב מכויל נכון מזהה יריב אסטרטגי — מייחס כוונות זדוניות להתנהגות תמימה. הסוכן מפתח אמונות שגויות בדבר כוונת פגיעה מכוונת, מגיב בהגנתיות כלפי שותף תמים, וסופג הפסדים מהותיים שסוכן מכויל היה נמנע מהם. כשל זה — יתר-מנטליזציה — מספק דין וחשבון חישובי פורמלי לפרנויה: לא כגירעון בקוגניציה חברתית, אלא כהפעלה פתולוגית מוגזמת. אנו מזהים את התנאים המשותפים שהם גם הכרחיים וגם מספיקים לכשל זה, מראים כי אותו מנגנון יכול להסביר חשיבה קונספירטיבית וחשדנות בהיקף רחב יותר, ומצביעים על קשרים למצבים קליניים ובהם אשליית רדיפה והפרעת אישיות גבולית.

התרומה השלישית שואלת האם קיים מודל שנמנע ממניפולציה, עם משאבים קוגניטיביים סופיים ומוגבלים. הפתרון טמון בהפרדה בין איתור למידול. במקום לנסות להתחכם למטעה מתוחכם — דבר שהוא בלתי אפשרי לוגית עבור סוכן בעל יכולת רקורסיבית מוגבלת — הקורבן יכול במקום זאת לנטר האם התנהגות המטעה עולה בקנה אחד עם כל מודל

עבודה זו נעשתה בהדרכתם של:

פרופ' ג'פרי רוזנשיין

פרופ' פיטר דיין

התפקיד של תיאוריה של התודעה במשחקי סכום כללי בייזיאנים

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

שם הסטודנט/ית

נתאי אלון

הוגש לסנט האוניברסיטה העברית בירושלים

מאי/2026

התפקיד של תיאוריה של התודעה במשחקי סכום כללי בייזיאנים

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

שם הסטודנט/ית

נתאי אלון

הוגש לסנט האוניברסיטה העברית בירושלים

מאי/2026