A (dis-)information theory of revealed and unrevealed preferences

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

[1]Hebrew University of Jerusalem [2]MPI for Biological Cybernetics [3]University of Tübingen

Abstract

In complex situations involving communication, agents might attempt to mask their intentions, essentially 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 strategically reinterpret or discard buyers' signals.

1 Introduction

Actions speak louder than words—sometimes enabling us to infer another person's beliefs and desires. Savvy speakers spin stories to fit their audience, like house buyers feigning disinterest to get a better deal; savvy listeners retaliate by ignoring them, like sellers sticking to their original prices. We employ information theory to analyze such signalling behavior in a simple two-agent task, showing how purely reward-maximizing agents endowed with a theory of mind (ToM) distort and re-interpret signals. We achieve this via the reinforcement learning (RL) framework of Interactive Partially Observable Markov Decision Processes (IPOMDP) [1].

IPOMDPs enable agents to plan through another agent's inference process via a theory of mind (Fig. 1). 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. A theory of mind such as this can be used for cooperation and deception [2, 3, 4, 5, 6], is hypothesized to underlie key parts of human economic and social cognition [7, 8, 9, 10, 11, 12, 13], and to arise already in childhood [14].

Information theory (IT) is a particularly helpful tool to understand the messages that get sent among agents in such complex interactions, particularly regarding deception [15, 16]. IT has also been used explicitly in multi-agent RL. Strouse et al. [17] showed that training the observed agent to minimize or maximize the Mutual Information (MI) between its action and goals improves (mis-)communication in cooperative and competitive games. In contrast, we show that this reduction emerges from reward maximization and theory of mind alone—without handcrafting any reward function.

To the AI community, our work demonstrates how complex signalling behavior can emerge via pure reward maximization [18]. To the cognitive science and economics communities, we highlight computations that show how humans might optimally act when signalling using theory of mind. Information theory provides a tool with which to understand deception and skepticism.

[*]Denotes joint first-authorship [†] Denotes joint last authorship.

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

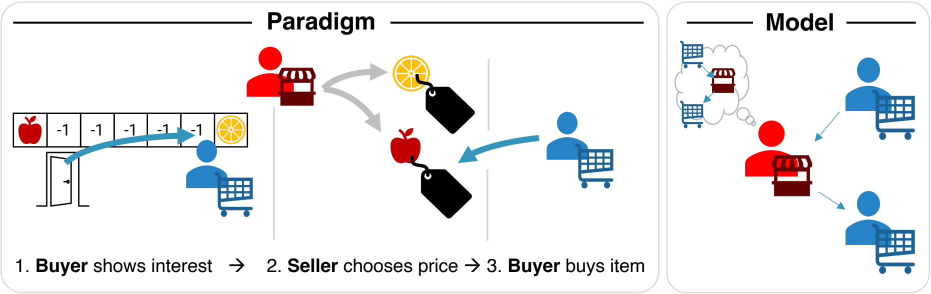
Paradigm
Model
-1
-1
-1
-1
-1
1. **Buyer** shows interest →
2. **Seller** chooses price → 3. **Buyer** buys item

Figure 1: Paradigm and Model

2 Paradigm, model, and agents

We model a buyer and a seller interacting over three stages (Fig. 1). Imagine a store owner offering items at stalls in various locations, and using observations of buyer behavior to set subsequent prices.

In the *first* stage, the buyer enters what can be thought of as a simple T-Maze with an apple and an orange located in opposite arms. The buyer incurs costs for travelling down one of the arms until it reaches and consumes one of the fruits, for which it duly receives a reward based on its preferences. Crucially, the buyer's decisions are observed by the seller [19] who in turn is aware of the buyer's travelling cost, but is *not* aware of its preferences.

In the *second* stage, the seller uses their observation of the buyer to set prices for a future possible purchase of one of these two items. In the *third* stage, the buyer purchases one of the items for the set price and then consumes it, again receiving a reward for this consumption. For illustrative purposes, we here restrict the preferences of oranges and apples to sum to 10. We impose the same restriction on the walking cost in the first stage and the prices the seller can set in the second stage.

We study different levels of sophistication of IPOMDP buyers and sellers arising from the recursive depth of their ToM. As the turns in this game alternate [13], the ToM levels of buyer and seller also alternate, starting from the simplest buyer, which we denote as ToM(-1). The ToM(-1) buyer independently decides based on cost/price and preferences at the first and second stages without taking into account the seller. The ToM(0) seller does Bayesian inverse reinforcement learning [20] on the buyer's first stage actions to infer the buyer's preferences and set its own prices. The ToM(1) buyer takes this inference process into account, and performs model-based planning through the seller's inverse reinforcement learning to optimize the sum of both first and second stage payoffs. The ToM(2) seller tries to defend against this hacking via higher-order inverse reinforcement learning, which the ToM(3) buyer again attempts to 'hack'. Unlike Camerer et al. [12], we assume a strict nesting, where each ToM-level uses only a model of the agent one step below on the ToM ladder. Furthermore, all buyers act in the same way in the final stage, purely maximizing their reward based on their preferences and the seller's prices, without taking into account any further step.

3 Results

We present the agents' policies resulting from this progressive, recursive modelling, focusing on the two signature decisions of buyer and seller: the buyer's first action, and the seller's corresponding price. In parallel, we will re-express the (mis-)information sent and received by the agents through the lens of information theory.

The ToM(-1) buyer acts naïvely, maximizing the utility of each stage separately. Fig. 2A duly shows the probabilities of choosing an apple. The apple is more likely to be chosen when it is closer (left of x-axis) and more preferred (top of y-axis). We can re-express this 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. This MI is shown in Fig. 2F—the action is generally informative, particularly when the fruits are nearly equidistant from the buyer.

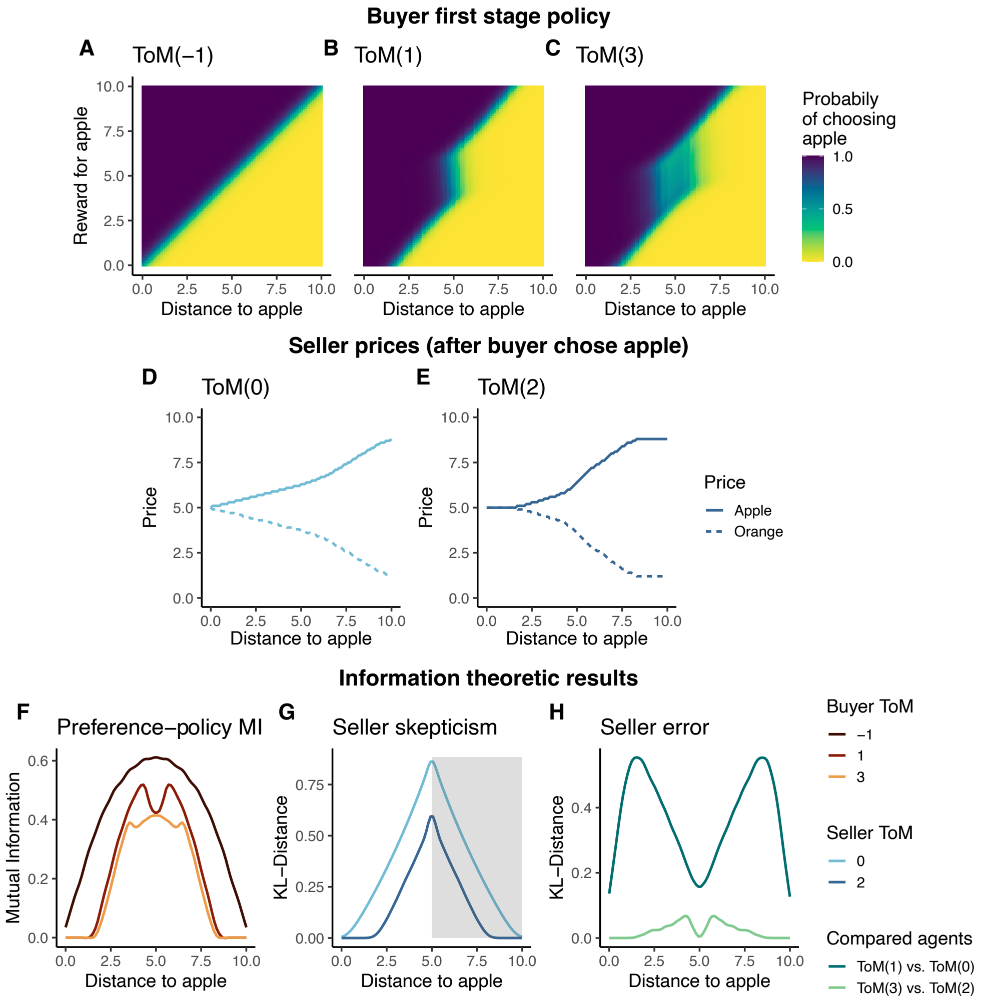

Figure 2: **(A-C)** Buyer policy in first stage as a function of the distance and the preference towards the apple shown by different ToM-levels. Note that we are using a soft policy with a temperature of $\beta = .5$. **(D-E)** Seller prices after a buyer chose the apple as a function of the distance to the apple, for the two different seller ToM-levels. **(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. **(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.

In turn the ToM(0) seller can translate what it knows about the ToM(-1) policy into prices via inverse reinforcement learning. Fig. 2D 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 prices match the distance walked. IT again lets us re-express this "trust" in the buyer's action in mathematical terms. Specifically, we can measure the strength of the seller's belief update via the Kullback-Leibler-divergence (KLD) between its (flat) prior and posterior. In Fig. 2G, we show the KLD for 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 step of the buyer as a signal.

Aiming to get the best price possible, the ToM(1) buyer attempts to hack this pricing scheme (Fig. 2) by playing what amounts to a gambit. This manipulation is, for example, evident in the lower left side of Fig. 2B. There, the buyer has little preference for the apple but is close to it. While the ToM(-1) will reach for the orange, the ToM(1) picks the apple. Crucially, this incurs the ToM(1) a small negative reward in the first stage, but will lead a ToM(0) seller to mis-price the apple, resulting in a greater overall reward.

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 (Fig. 2F). Furthermore, 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 [15] 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. This belief discrepancy is shown in Fig. 2H, highlighting the large discrepancies.

Having access to this ToM(1) buyer model, the ToM(2) seller becomes skeptical about the buyer's actions, and adjusts its pricing appropriately. We show this pricing in Fig. 2B. When the maze setting enables the ToM(1)'s bluff (for example, when the apple costs are 1), observing an apple selection provides the seller with no information about the buyer's preferences (compare the relevant policy of the ToM(1) agent 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 Fig. 2G that shows how the ToM(2) seller's belief about the buyer is affected less, or, when the items are closer, not at all, by the buyer's actions).

The ToM(3) attempts to maneuver around this skeptical pricing to achieve the best overall reward. However, it is essentially cornered and can only attempt minimal ruses in a few possible game settings, particularly when the items are roughly equidistant (see Fig. 2C). 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 outmaneuver 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 (Fig. 2F). 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 (Fig. 2H). Note that the ToM(2) dissimilarity increases in regions where the ToM(0) dissimilarity decreases, showing the ToM(3)'s attempts at deception.

4 Discussion and future work

Our work shows how purely reward-maximizing agents can appear to engage in complex information-theoretic signalling behavior. Crucially, unlike Strouse et al. [17] we do this in the absence of any hand-crafted value function and only rely on theory of mind and planning. More specifically, we show how senders can purposefully reduce the informativeness of their actions and target a receiver's inference process. Equally, savvy receivers can become skeptical and selectively ignore or reinterpret a sender's signal.

Our work is relevant for the study of social cognition in artificial [21, 22] and biological systems. For example, it adds a reinforcement learning and information theoretic perspective to Goodhart's law [23], which states that people tend to try to game statistical regularities used by authorities (e.g., the government or a retailer) for control purposes (e.g., taxation or dynamic pricing). There are many avenues for further enquiry. For example, how closely do humans [24, 25], or other animals [26, 27], actually follow our theoretic analyses? Furthermore, an expanded model might allow us to ask whether rational actors would choose to pay to observe, or for being observed, or how savvier sellers might act to shape an interaction to be especially informative.

Acknowledgements

We would like to thank Rahul Bhui and Stefan Bucher for helpful comments. This research has been partly funded by Israel Science Foundation grant #1340/18 (NA; JR), by the Max Planck Society (NA, LS, PD) and the Humboldt Foundation (PD).

References

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

[2] 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. Springer International Publishing, 2021. ISBN 978-3-030-91779-1. doi: 10.1007/978-3-030-91779-1_5.

[3] Leander Vignero. Updating on biased probabilistic testimony. *Erkenntnis*, pages 1–24, 2022.

[4] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016. ISSN 1364-6613. doi: 10.1016/j.tics.2016.08.005. URL `https://www.sciencedirect.com/science/article/pii/S136466131630122X`.

[5] Michael Franke, Giulio Dulcinati, and Nausicaa Pouscoulous. Strategies of deception: Under-informativity, uninformativity, and lies—misleading with different kinds of implicature. *Topics in Cognitive Science*, 12(2):583–607, 2020.

[6] Lauren A Oey, Adena Schachner, and Edward Vul. Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 2022.

[7] Mark K. Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 2022. ISSN 1364-6613. doi: 10.1016/j.tics.2022.08.003. URL `https://www.sciencedirect.com/science/article/pii/S1364661322001851`.

[8] 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.

[9] Ting Xiang, Debajyoti Ray, Terry Lohrenz, Peter Dayan, and P Read Montague. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, 8(12):e1002841, 2012.

[10] Benedetto De Martino, John P O'Doherty, Debajyoti Ray, Peter Bossaerts, and Colin Camerer. In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6):1222–1231, 2013.

[11] Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488, 2020. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2020.107488. URL `https://www.sciencedirect.com/science/article/pii/S0028393220301597`.

[12] Colin F Camerer, Teck-Hua Ho, and Juin Kuan Chong. A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3:157–162, 2015.

[13] 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, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004254. URL `https://dx.plos.org/10.1371/journal.pcbi.1004254`.

[14] 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, 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.05.011. URL `https://linkinghub.elsevier.com/retrieve/pii/S1364661316300535`.

[15] 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, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207383. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207383`. Publisher: Public Library of Science.

[16] Noga Zaslavsky, Jennifer Hu, and Roger P Levy. A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*, 2020.

[17] DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, and David J Schwab. Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/1ef03ed0cd5863c550128836b28ec3e9-Abstract.html`.

[18] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103535. URL `https://www.sciencedirect.com/science/article/pii/S0004370221000862`.

[19] Shuwa Miura and Shlomo Zilberstein. A unifying framework for observer-aware planning and its complexity. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 610–620. PMLR, 27–30 Jul 2021. URL `https://proceedings.mlr.press/v161/miura21a.html`.

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

[21] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[22] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

[23] Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pages 91–121. Springer, 1984.

[24] Keith Ransom, Wouter Voorspoels, Amy Perfors, and Daniel Navarro. A cognitive analysis of deception without lying. Cognitive Science Society, 2017.

[25] Samuel A Barnett, Robert D Hawkins, and Thomas L Griffiths. A pragmatic account of the weak evidence effect. *arXiv preprint arXiv:2112.03799*, 2021.

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

[27] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.