Theory of Mind in Partially Observed, Mixed-Motive Games

Nitay Alon

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

Abstract

Theory of Mind (ToM) enables agents to model others' mental states, but in mixed-motive games, this capacity can lead to deceptive behaviour and alignment risks. My research investigates how ToM affects strategic behaviour in partially observed games, contributing: (1) a formal model of ToM-driven manipulation in a preference elicitation task, (2) evidence that excessive ToM leads to paranoid-like overmentalisation, and (3) the Aleph-IPOMDP model, a framework for multi-agent systems that balances ToM reasoning with game-theoretic principles to prevent manipulation, deterring capable agents from deceiving. My work contributes to the understanding of deceptive AI, overcoming deception in multi-agent systems and applications to computational model of human cognition.

Introduction and Motivation

Theory of Mind is one of the hallmarks of human cognition (Premack and Woodruff 1978). This trait endows us with the capacity to infer and simulate others' mental processes like beliefs, desires and intentions. While effective for cooperative tasks such as multi-agent communication, this capability can also facilitate the learning of malevolent behaviour. For example, children learn to deceive other kids, once trained with ToM (Ding et al. 2018). My research aims to identify the role of ToM in learning deceptive behaviour and counter deceptive behaviour. Specifically, this involves modelling how ToM agents interact in non-cooperative games with the purpose of mapping the relation between situations and emergence of deceptive behaviour. With growing interest in augmenting AI with ToM my work seeks to detect and prevent AI-safety issues such as deceptive AI.

I address this through several interconnected sub-projects. First, I explored how agents learn to manipulate others, hiding their intent from other agents for strategic gain (Alon et al. 2023). One key finding from this work is the emergence of skepticism as a function of the cognitive hierarchy of the agent. Building on this observation, I explored the rise of paranoid-like behaviour in mixed-motive games (Alon et al. 2024). This work illustrates that while high ToM levels are beneficial for designing and detecting deceptive behaviour, it can lead to overmentalisation, harming the agent. To overcome this issue, I designed a new framework, the ℵ-IPOMDP (Alon et al. 2025) in which agents shift between using ToM and Game-Theoretic principles to avoid both over and under mentalisation. Overall, my previous work contributes to the understanding of the interaction between environment (mixed-motive games) and emergence of deceptive behaviour and proposes potential remedies to overcome this issue.

Background and Relevant Work

Previous studies on strategic ToM behaviour used well designed situations to illustrate how ToM agents use it to gain advantage. For example, negotiation (de Weerd, Verbrugge, and Verheij 2017), in which agents use their ToM for both goal inference as well as for strategic planning. I extended this approach to other games to identify which game structures incentivize deceptive behaviour. I showed how deception arises in partially observed games, where agents can manipulate the perception of others.

The Interactive-POMDP [IPOMDP;(Gmytrasiewicz and Doshi 2005)] framework allows to represent ToM agents in a symbolic manner, combining POMDP inference with nested opponent reasoning. While being beneficial for modelling social interactions, it lacks the capacity to model situations where an agent interacts with an unmodeled opponent (one with higher ToM level), a gap the ℵ-IPOMDP (Alon et al. 2025) solves by augmenting this framework.

Theoretical models of ToM and deceptive AI presented by (Sarkadi 2023). My work focuses on modelling environmental effects on agents' tendency to deceive. While previous work suggested that deceptive behaviour should stimulate the rise of a "cognitive" arms-race (Sarkadi 2023), my work shows a different pattern, following the work of (Piazza, Behzadan, and Sarkadi 2023) I showed that ToM can lead to self-harming behaviour due to overmentalisation.

Current Progress and Results

Published work In my work on strategic dis-information (Alon et al. 2023) I developed an economic model of buyer-seller interaction. In this model ToM-endowed agents, modelled using the IPOMDP framework, use their ToM to predict how their counterpart will infer their intent (item preferences) from observed behaviour. This leads to a variety of

deceptive and counter-deceptive behaviours. For example, a ToM endowed buyer feigns interest in the less preferred item to falsely signal preference with the intent of causing the seller to disregard observed behaviour. This causes a ToM-endowed seller to discount observed behaviour by ignoring it, measured as reduced KL divergence between prior and posterior beliefs.

Given the distrustful behaviour learnt by high ToM agents, my work on overmentalisation (Alon et al. 2024) suggested that *high* levels of ToM may cause paranoid behaviour, contradicting the commonly assumed paradigm that people with paranoid behaviour lack or have dysfunctional ToM. This work showed that while high level of ToM robustifies the agent's ability to resist manipulation by "seeing through the bluff', it also makes agents susceptible to ascribing intentionality to random or benign behaviour. Combining the findings of both papers, I proposed the א-IPOMDP (Alon et al. 2025) (under review) a computational model combining the benefits of ToM for multi-agent planning with anomaly detection components to identify when its predictions about others fail, signalling that its ToM is miscalibrated. Once identified, the agent switches to a precompouted policy (MinMax in zero sum and grim trigger in mixed-motive games) to avoid being taken advantage of. I illustrated how this model reduces deceptive behaviour by serving as a credible threat for agents with higher ToM level than the א agent, improving the א agent's utility compared to the baseline case.

Future work My work showed how ToM can be used for deception, and in the process it uncovered two major issues with ToM: *having the correct model of the other* and *adapting reasoning level to the opponent and situation*. Considering the complexity of human society, where interactions with unknown others in various social interactions are omnipresent, it is unlikely that we form a computational model of each person we interact with, nor do we engage in computationally demanding recursive reasoning for no good reason. My current research aims to understand *how AI agents can effectively adapt their ToM without these pitfalls*.

Solving these problems is fundamental to artificial ToM (A-ToM). I aim to model how agents, Humans and AI alike, adjust their mentalisation to match both the *complexity* of others and the *nature* of the interaction, while being resource (and outcome) efficient. While moving from strict k-level model to the broader Cognitive-Hierarchy (CH) (Camerer, Ho, and Chong 2004) is possible, it is still computationally demanding.

With the adapted model, my next goal is to identify the *triggers* that cause shifts to reasoning levels and link these cues to maladaptive ToM. For example, if agents tend to overmentalise due to situational misinterpretation (mistaking cooperative task for a competitive one). Using this information, I plan to *regularize* the mentalisation by incorporating ideas from self-regulation theory and build a meta-learning paradigm, that monitors and regulates mentalisation levels. The end result is a single model of ToM, capable not only of reproducing human ToM, but also augment current AI agents to mitigate Human-AI alignment issues.

Contribution My thesis contributes to both AI and Cognitive Science in multiple ways. **Methodological**: The א-IPOMDP framework proposes a robust model for deception mitigation, allowing agents to cope with unmodeled or oversophisticated others making them somewhat resistant to manipulation with improved social welfare results. **Theoretical**: A formal characterization of how ToM affects strategic behaviour in mixed-motive games, with detailed analysis of each ToM-level and the associated strategies. My work presents an interpretable and measurable model of deception using Information Theoretic metrics, allowing us to map from behaviour (action) to intent (manipulation) and vice versa. **Empirical:** By illustrating the risks associated with overmentalisation, my work contributes to both AI-Safety community and the Computational Psychiatry domain, as we showcase how ToM may lead those bestowed with it astray — leading to detrimental individual and social results.

References

Alon, N.; Barnby, J. M.; Sarkadi, S.; Schulz, L.; Rosenschein, J. S.; and Dayan, P. 2025. Detecting and Deterring Manipulation in a Cognitive Hierarchy. ArXiv:2405.01870 [cs].

Alon, N.; Schulz, L.; Bell, V.; Moutoussis, M.; Dayan, P.; and Barnby, J. M. 2024. (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*, 8(1): 159–177.

Alon, N.; Schulz, L.; Rosenschein, J. S.; and Dayan, P. 2023. A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind*, 7: 608–624.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A Cognitive Hierarchy Model of Games*. *The Quarterly Journal of Economics*, 119(3): 861–898.

de Weerd, H.; Verbrugge, R.; and Verheij, B. 2017. 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.

Ding, X. P.; Heyman, G. D.; Sai, L.; Yuan, F.; Winkielman, P.; Fu, G.; and Lee, K. 2018. Learning to deceive has cognitive benefits. *Journal of Experimental Child Psychology*, 176: 26–38.

Gmytrasiewicz, P. J.; and Doshi, P. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24: 49–79.

Piazza, N.; Behzadan, V.; and Sarkadi, S. 2023. Limitations of Theory of Mind Defenses against Deception in Multi-Agent Systems. preprint, In Review.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526. Publisher: Cambridge University Press.

Sarkadi, S. 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.