Cutting Corners in Theory of Mind*

Marion Roth[1,†], Stacy Marsella[1,2] and Lawrence Barsalou[1]

[1]University of Glasgow, University Avenue, Glasgow, G12 8QQ, United Kingdom
[2]Northeastern University, 360 Huntington Ave, Boston, MA 02115, USA

Abstract

Theory of Mind (ToM) research is concerned with the fundamental social ability of cognitively representing other people's mental states. In the context of designing intuitive, human-like, and collaborative artificial intelligence (AI), ToM is becoming increasingly relevant to AI researchers. However, there is a missing link between the theoretical and empirical approaches to ToM in Psychology and the applied but theory- and evidence-distant work in Computer Science. This work aims to connect the understandings and aims of both fields. A dual-process theory driven framework was adopted, distinguishing between higher-level and lower-level processes. Participants saw a virtual agent's movement in a maze with either a full egocentric cue, a partial egocentric cue, or no egocentric cue. They were asked to predict the agent's next actions and give verbal explanations for their predictions. Results show a strong sensitivity to egocentric information. Moreover, the qualitative responses show large individual differences not reflected in the recordings of participants' decisions. The mechanisms contributing to efficient and flexible ToM and their implications for applied ToM in AI are considered and discussed. The value of qualitative research in this subject area is highlighted.

Keywords

Theory of Mind, Artificial Intelligence, Efficiency, Heuristics, Dual Process Theory

1. Introduction

Theory of Mind (ToM) is the ability to reason about other people's mental states [1]. It is also commonly referred to as perspective-taking or mentalising and has a central role in human social interaction. Much of human communication is not explicitly verbalised [2]. Instead, people use their abilities to infer others' beliefs, intentions, and feelings to make sense of social events.

Since 1978, when Premack and Woodruff coined the term, research has greatly evolved. There are various theories explaining this ability and the application of the concept has vastly expanded in recent years. There is now research on the development, social relevance, influencing factors, cultural differences, individual differences, and neuroscience of ToM.

AAAI 2022 FALL SYMPOSIUM SERIES, Thinking Fast and Slow and Other Cognitive Theories in AI, November 17-19, Westin Arlington Gateway in Arlington, Virginia, USA

*This work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1

†Corresponding author.

m.roth.1@research.gla.ac.uk (M. Roth); Stacy.Marsella@glasgow.ac.uk (S. Marsella); Lawrence.Barsalou@glasgow.ac.uk (L. Barsalou)

https://stacymarsella.org (S. Marsella); http://barsaloulab.org (L. Barsalou)

0000-0002-0345-9226 (M. Roth); 0000-0002-5711-7934 (S. Marsella); 0000-0002-1232-3152 (L. Barsalou)

Beyond psychological research, Theory of Mind has increasingly been studied in the field of Artificial Intelligence (AI). In the context of the development of AI that fits into human society, the demand for social intelligence in artificial agent grows [3]. Currently, the implementation of typical human traits in artificial intelligence is exploding as a research field.

However, there is a fundamental difference in perspective on intelligence and behaviour from psychological and computational research. The two fields are highly complementary and have the potential to greatly inform each other but there is limited exchange of concepts across them. For example, there is evidence that humans use their ToM abilities very selectively [4]. Computational models of ToM, in contrast, often are built on the assumption that we create rich models of people all the time [5]. In a world as complex as ours, this strategy would be impossible to realise. The heuristics and shortcuts the brain uses to perform ToM as flexibly as psychologists have observed are largely understudied and overlooked in the field of Social AI.

In this paper we discuss ToM at the intersection of Psychology and Computer Science. We illustrate the complexity of ToM and highlight the challenges in generating computational models. We then present an experimental study exploring human ToM use to understand the heuristic mechanisms underlying ToM, towards developing a more efficient computational model. The results suggest a strong influence of egocentric information in predicting others' behaviour. The possible ramifications for computational models are discussed.

2. Background

This section considers the question of Theory of Mind reasoning from both a psychological perspective and a computational standpoint.

2.1. Psychological perspectives

The field of ToM is vast. Most work is developmental research studying if, when, and how ToM develops in young children. The focus of the present work, however, is on adults, assuming a typically functioning ToM system. The emergence of different theories of ToM have led to a variety of understandings of what the ability entails. Generally, ToM is understood as the ability to reason about others' mental states. However, where this ability starts and ends is not clearly defined [6]. Furthermore, it has been argued that ToM tasks measure different aspects of ToM [7].

Ultimately, ToM is an umbrella term for different phenomena. A person's model of another human could consist of any of, and is not limited to, the following:

- explaining a person's behaviour by modelling their goals and/or moral values
- predicting a person's behaviour by modelling their goals and/or moral values
- modelling a person's beliefs by interpreting their behaviour
- inferring a person's intentions by considering one's own experience
- deciding on one's own behaviour by modelling another person

Moreover, mental models can vary greatly in accuracy and depth. The ability to use ToM is also characterised by a large variance in ToM use across individuals and situations [8].

Psychological research suggests that humans do not always infer others' mental states, even when it would be helpful to do so [4]. Studies of cultural differences in ToM suggest an element of individual learning and experience in shaping those tendencies [9]. Furthermore, there is evidence that the activity of another person, as opposed to them being passive, facilitates perspective-taking [10]. Moreover, the nature of the other appears to be relevant for ToM use, i.e. whether it is a human or an inanimate object [11]. Research also suggests that mood and emotion are related to the extent to which an individual may be egocentric in their perspective [12].

Furthermore, researchers have distinguished between different types of ToM. The distinction between explicit and implicit ToM [13], for example, is in line with other dual-system approaches to psychological phenomena [14, 15]. Proponents of these accounts generally distinguish between a fast system and a slow system. The fast system uses implicit and automatic processing, based on previously established beliefs and reaction patterns, which we are not consciously aware of. It is characterised by heuristics and biases, which make it fast but also difficult to intervene with or execute direct control over. The slow system is believed to be deliberate, operating within the scope of our conscious awareness.

Apperly and Butterfill [13] strongly argue in favour of a dual-system account of mindreading, suggesting that the fast and implicit component of Theory of Mind develops early, whereas the explicit, slow, and more cognitively demanding element develops only later-on. They propose that young children fail at the explicit but not the implicit mind-reading skills [16]. Some ToM research focusses primarily on explicit ToM and some primarily on implicit ToM, whereas other researchers discount the distinction altogether. Burge [17], for example, argues that there are very powerful other explanations for the evidence suggesting Theory of Mind in very young children, which would not require mind-reading, such as learning based on associations or reinforcement.

The focus of the present work is on explicit ToM. This is not because the work argues against implicit ToM requires perceptual measures [18], which are not applied here. Instead, this work specialises on the explicit content of mental models and the computations by which it is selected. Subsequent sections of this paper will, however, suggest a more complex relationship between implicit and explicit processes. Specifically, implicit processes may shape a person's explicit interpretations of behaviour.

2.2. Computational perspectives

In response to increasing demands of AI to complete social tasks [19], especially in collaboration with other humans, multi-agent research has explored a range of techniques for agents to model other agents. These include but are not limited to the use of action observations to reconstruct action strategies, modelling group relations to predict agents' decision making, or using hierarchical action descriptions to determine plans and goals. See Albrecht & Stone [20] for a detailed review. Here we briefly cover the work relevant for the current study.

With growing insight that the brain operates like a "prediction machine" [21], researchers have increasingly used Bayesian approaches to model ToM. Baker and colleagues [5], for example, developed the Bayesian Theory of Mind (BToM), which is based on probabilistic backward inferences. The model computes which mental state is most likely given the observed

behaviour. BToM accurately represents human responses on different tasks. However, the domains are very restricted by the design of the study. Instead of forcing a flexible reaction to a rich environment or situation, the search space at hand is tailored to the model from the start.

Models of others have been described as higher-order beliefs [22], i.e. beliefs about beliefs, with varying levels of depth. Computationally they have been represented as nested beliefs with a recursive structure [23]. Applying this principle, for example, PsychSim [24, 25] is a system which models multiagent sequential decision-making tasks. IPOMDPs (interactive partially observable Markov decision processes) generate forward inferences to predict others' behaviours.

Especially when several levels of depth are involved, computing all possible behaviours a person may engage in (considering several steps of projection into the future and a variety of possible events or behaviour by others), to then choose the most likely option, can be extremely intractable. Considering that humans operate in very complex environments, it seems highly unlikely that a person could realise this mechanism explicitly, given short-term memory constraints, at least to the full extent of searching through all possible mental states. This therefore poses the question when exactly human ToM is performed and to what degree. What are the computational shortcuts human brains take to speed up processing?

3. Rationale

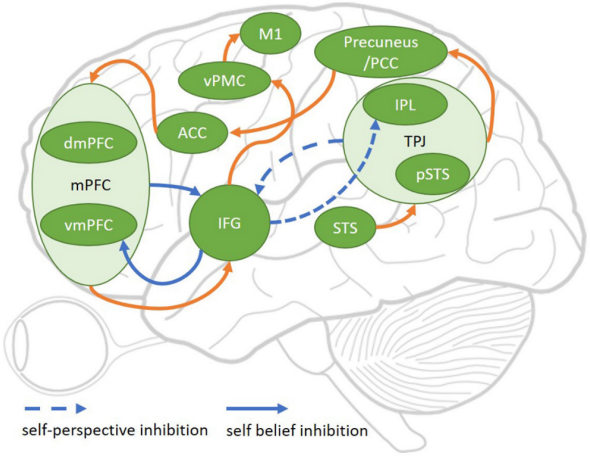
In order to create a realisable computational model of ToM it is required to work around this problem of tractability. Towards that end, we can ask what decides when humans use ToM, how they use it, as well as when they do not use it. What features of the situation, experience and social interaction determine the activation of human mentalising abilities?

The potentially ToM-relevant feature illuminated in this work is the presence and timing of an egocentric cue. Zeng et al. [26] present a neuroscience-informed model, recognising the different brain areas that have been shown to be involved in ToM. They distinguish between perspective-type and belief-type information impacting on ToM use. Thereby, the perspective is an early element in mentalising, composed of information from lower-level brain areas, whereas beliefs are influential later-on, and the product of higher-level processing (Figure 1).

This model is consistent with dual process approaches to cognition, distinguishing between more automatic, fast, less controlled processes and more deliberate, slow, and conscious processes, and with models which distinguish between implicit and explicit ToM.

4. Hypotheses

Zeng et al. [26] distinguish between the self-experience learning pathway, the motivation understanding pathway, the reasoning about one's own belief pathway, and the reasoning about other people's belief pathway. Based on their model, the present study seeks to distinguish between subjects use of previously established higher-level ToM beliefs about the observed agent knowledge and low-level egocentric cues available to the participant. Specifically, participants are asked to predict the behavior of agent looking for a misplaced book that subjects are told is hidden from the agent. The experiment manipulates whether and when the subject gets a cue



Brain Area	Function in the computational model
STS	visually encodes biological motion
TPJ	distinguishes self and others store relevant stimuli
IPL	stores self-relevant stimuli
pSTS	stores other-relevant stimuli
Precuneus	infers visual access
PCC	sends information
ACC	belief reasoning
mPFC	stores beliefs
vmPFC	stores self-belief
dmPFC	stores other's belief
IFG	self-perspective inhibition self-belief inhibition
vPMC	encodes kinematics
M1	motion execution

Figure 1: The Brain-ToM model (including major functional brain areas, pathways, and their interactions) by Zeng et al. (2020).

as to the book's location. Thus the present study distinguishes between previously established beliefs and momentarily available egocentric cues. The following hypotheses were proposed:

H1: With no visual cue present, there is no egocentric bias (belief unbiased and perspective unbiased).

H2: With an early and partial visual cue present (belief biased but perspective unbiased), there is no egocentric bias.

H3: With a full visual cue present, there is an egocentric bias towards the target direction (belief biased and perspective biased).

The results of this study will show whether the distinction between lower-level and higher-level influences is helpful in modelling ToM.

5. Methods

5.1. Participants

This study was approved by the ethics board of the Psychology department at the University of Glasgow. 60 participants were recruited with the online platform Prolific. They were from the United Kingdom with a 95% approval rate on Prolific and at least 10 previously completed studies. 21(35%) of the participants were male, 38(63.33%) female, and 1(1.67%) other. The mean age was 39.43(SD=14.90) years with a minimum of 19 and a maximum of 67.

5.2. Apparatus

The maze videos used in this study are created with the tool used by Pöppel et al. [27]. A virtual agent moved along a pre-determined trajectory through a 2D space, searching for a book (see figures 2-4). Mazes were designed to rule out asymmetries in orientation as confounding variables. The target direction (egocentric cue present) was kept constant as the red option.

Videos were 10 seconds long, with the first static frame with the agent at its starting point and instructions (and cue if applicable) shown for 3 seconds, the agent remaining at the starting point without instructions (and cue in the second condition) for 1 second, and finally the agent moving around the space for 6 seconds.

5.3. Design

The study had a 1x3 between-groups design with the independent variable cue visibility (no cue vs partial cue vs full cue). Figures 2-4 show snapshots of the different conditions. The dependent variable direction was measured on a nominal scale with the categories red, black, and ”I don’t know”. Another dependent variable explanation was measured qualitatively by asking each participant why they chose their respective response.

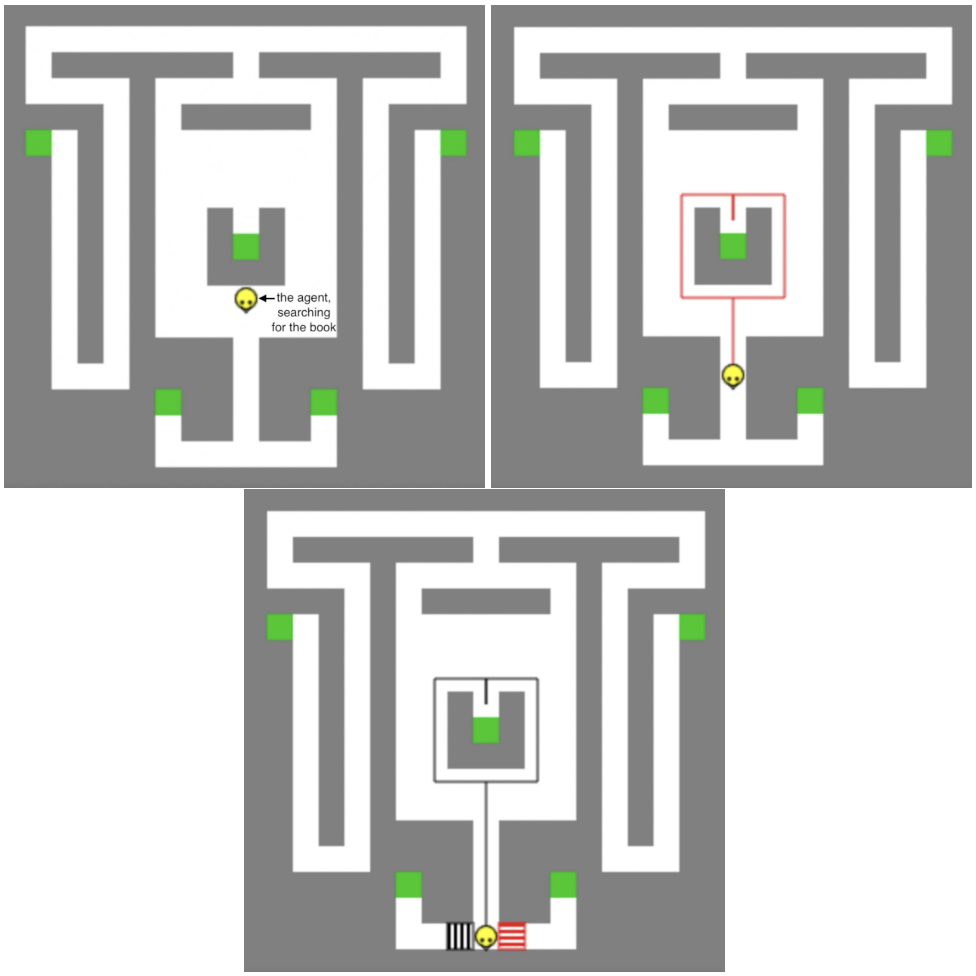


Figure 2: Condition 1 - no cue (participants do not see where the book really is).

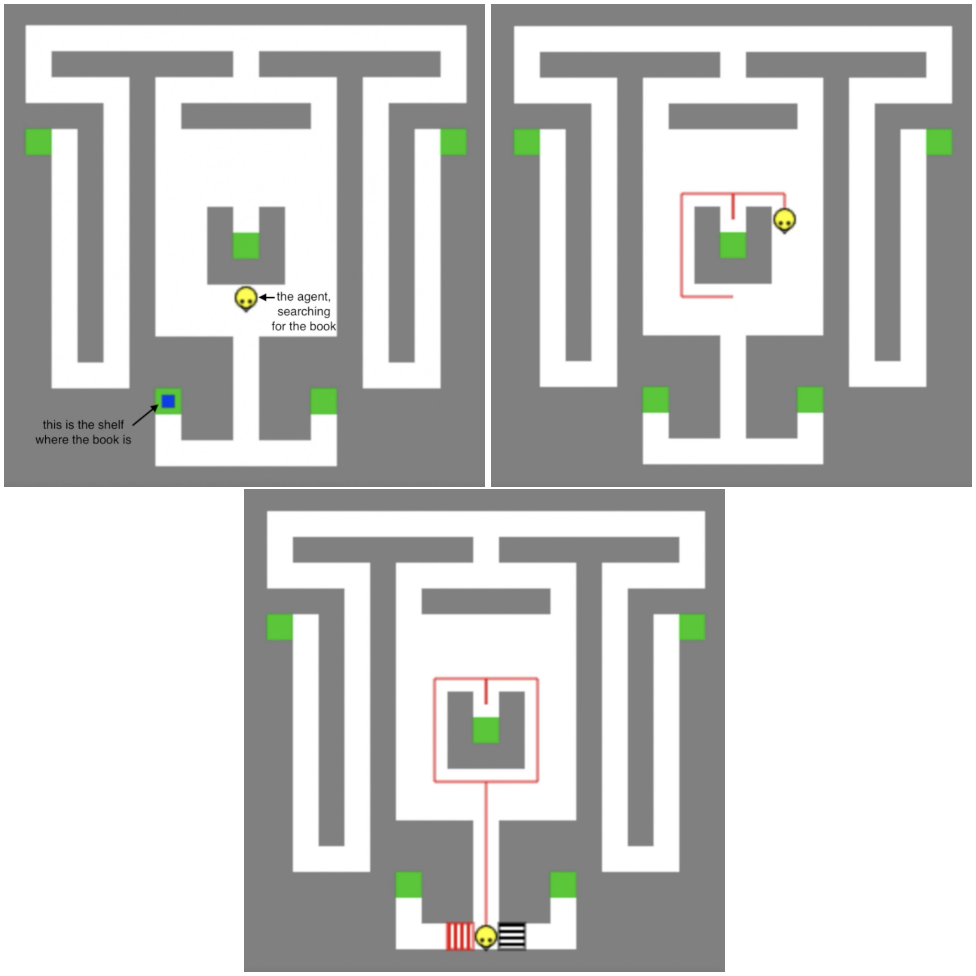


Figure 3: Condition 2 - partial cue (participants initially see where the book really is but then it disappears, and the cue is not visible at the time of the prediction).

5.4. Materials

The target stimulus included a short video of the mazes, with the cue visibility according to the condition participants were in. They were told that the agent was looking for the book and saw it move through the maze until it stopped at the bottom. At this junction where it turned either left or right, the participant was asked what they thought where the agent would go next. Thereby, one direction was marked in red, the other in black. Participants selected their answer from three choices: black, red, or "I don't know" (Figure 5).

The potentially confounding variables maze orientation (left vs right) and colour (black & red) were randomised for all conditions.

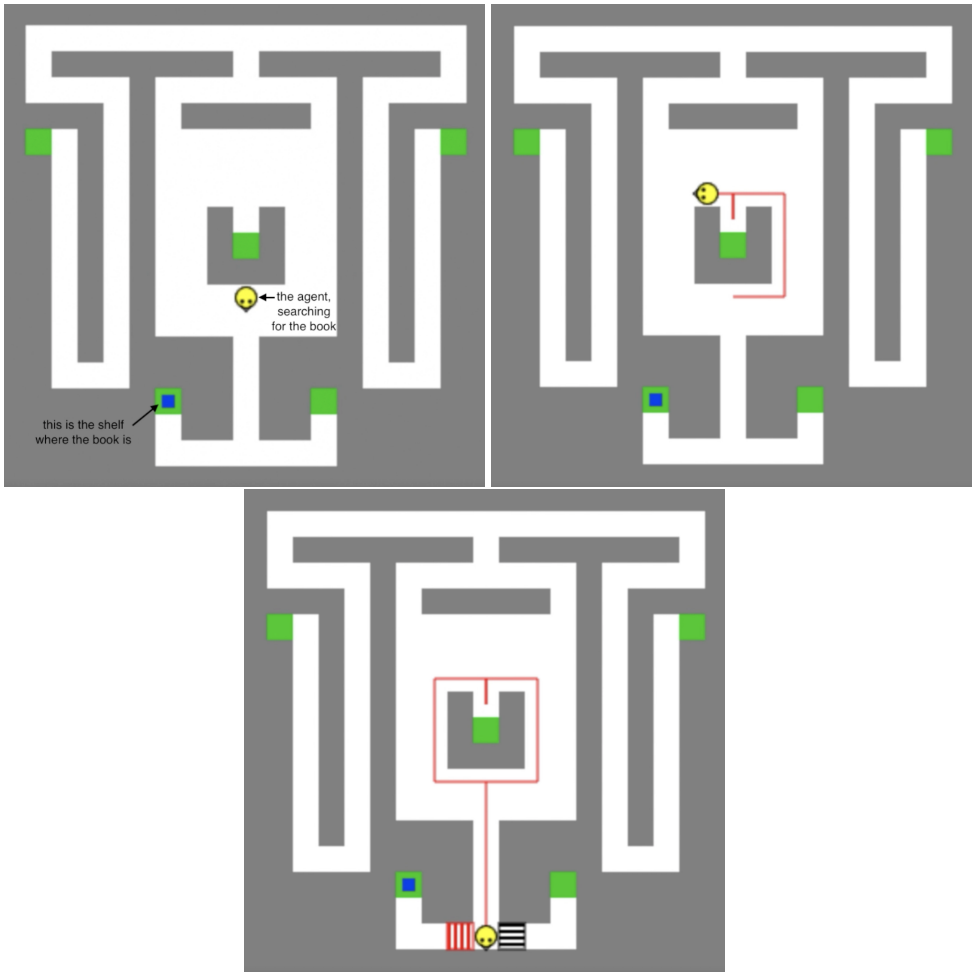


Figure 4: Condition 3 - full cue (participants see where the book really is from the start to the end, including the time of the prediction).

Figure 5: Answer choices.

6. Results

None of the demographics were found to significantly affect participant responses.

6.1. Quantitative Responses

Figure 6 shows which direction participants predicted the agent to go. When no cue was present, they equally chose between red and black, whereas the cue invoked favouring the red direction, regardless of when it was shown. A Chi-Square analysis was conducted and shows that the

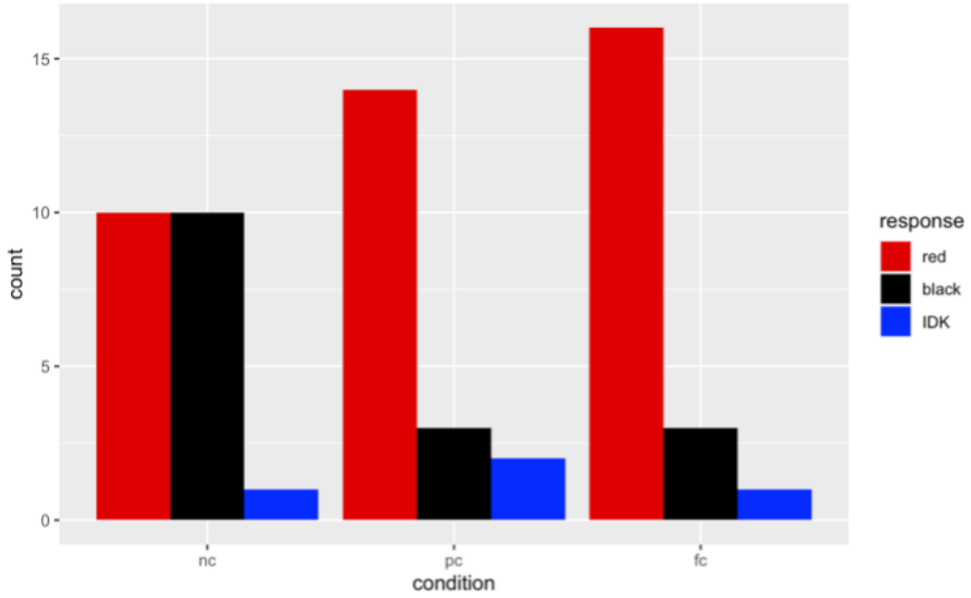


Figure 6: Counts of direction responses (black, red, I don't know) by cue visibility (no cue, partial cue, full cue).

difference between the three groups is not significant, X2(4, N=60) = 7.76, p = 0.101.

6.2. Exploratory Analysis of Qualitative Responses

We asked participants why they chose the answer they selected. As depicted in the framework by Zeng et al. [26], qualitative responses were analysed with regards to the distinctions between faster, more intuitive lower-level (perspective) pathways vs slower, more rational higher-level (belief) pathways, and self (egocentric) vs other (altercentric).

1. EP = Egocentric Perspective (self-experience learning, i.e. own feeling, intuition, etc.)
   "it just felt right"
   "I always read and choose left to right"
2. AP = Altercentric Perspective (motivation understanding, i.e. prediction of other's actions based on the experience of their previous actions)
   "Because he kept turning right on the search"
   "it always went to its left"
3. EB = Egocentric Belief (reasoning about one's own belief)
   "closest to the book"
   "It was where the book was"

4. AB = Altercentric Belief (reasoning about other people's belief)
   "There is no way to know"
   "because it's a 50/50 chance, I don't know"

We analysed whether participants in the different cue visibility conditions chose different strategies to predict the agent's next move. The Chi-Square test shows that they did, X2(6, N=60) = 12.84, p = 0.046. There was no significant difference by cue visibility in the source of information (belief vs perspective) for participants' prediction of where the agent would go, X2(2, N=60) = 0.45, p = 0.798. (Figure 7). Interestingly, perspective-based reasoning was more common than belief-based reasoning in all three conditions. Participants did, however, differ

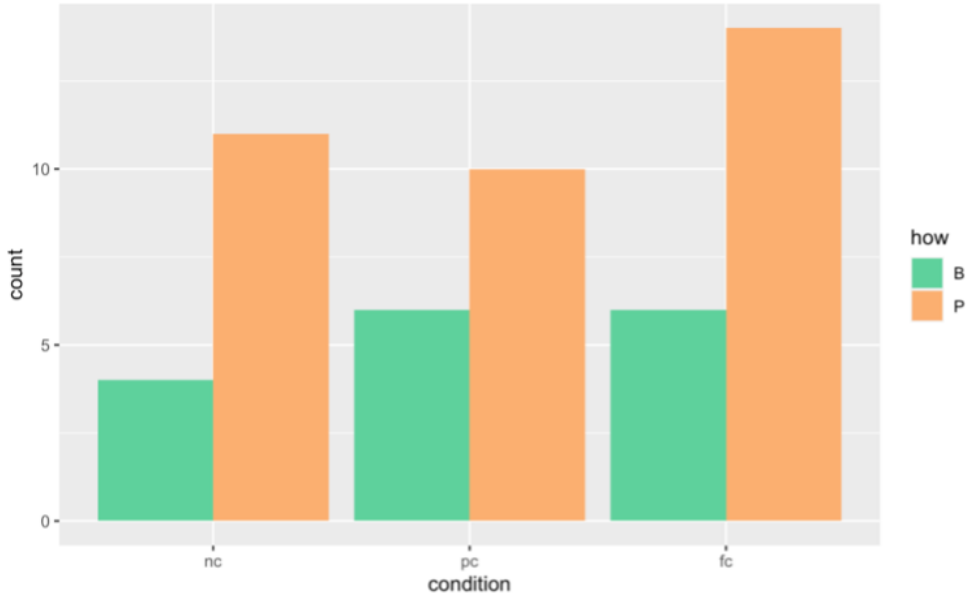


Figure 7: Counts of qualitative responses based on the source of information considered (belief vs perspective).

significantly in what viewpoint they took depending on whether they saw a cue or not, X2(2, N=60) = 10.85, p = 0.004 (Figure 8). When no cue was visible, participants largely considered the agent's perspective (A) but when a cue was there, the own perspective (E) dominated, regardless of whether the cue was only shown early or throughout the video.

7. Discussion

7.1. Quantitative Responses

It was hypothesised that there is an egocentric bias with a visual cue fully present (H3) and no egocentric bias when the cue is only partially present (H2) or absent (H1). The descriptive statistics suggest an egocentric bias in both the early cue and full cue condition and not in the no cue condition (Figure 6). Statistical analyses do not support the significance of these

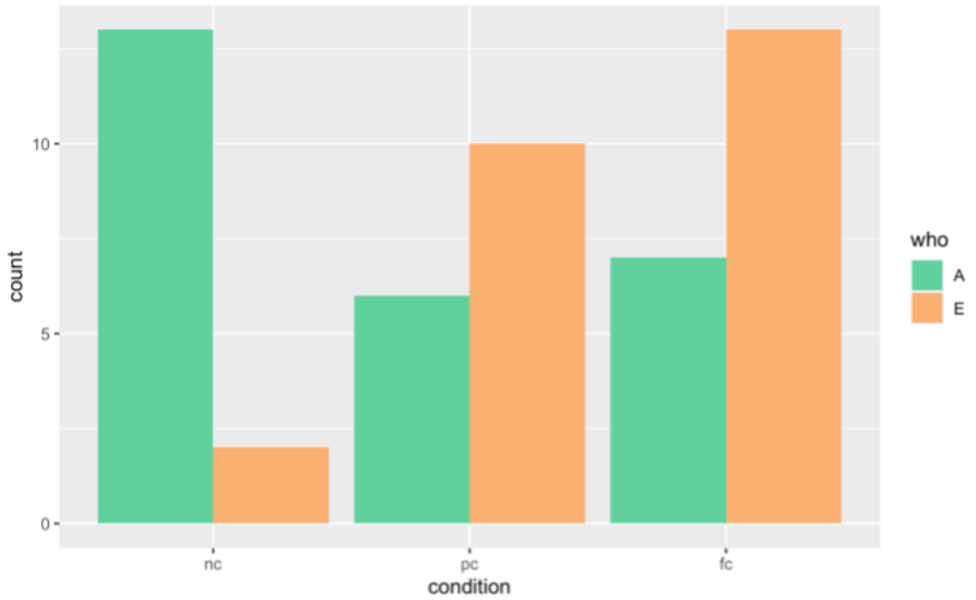


Figure 8: Counts of qualitative responses based on the viewpoint taken (own/egocentric vs other/altercentric).

differences. However, the result is close to significance and possibly underpowered with only 20 participants in each condition. We therefore recommend to follow this study up with a larger sample size. The result did not reach significance, but the trends are consistent with the phenomenon shown in previous literature that egocentric information is very dominant in social interactions (Keysar et al., 2003).

7.2. Qualitative Responses

Analyses of participants' verbal statements show that whether participants based their choices on belief-based or perspective-based explanations did not differ across conditions (Figure 7). Perspective-based reasoning was more common than belief-based reasoning in all three conditions. Zeng et al. [26] suggest that this pathway is faster, which suggests that it may be favoured by the brain to save costs and energy.

However, the dual process approach applied here is not without its criticism. While the framework has gained a lot of popularity, researchers have also highlighted its limitations. Evidence suggests that many psychological findings require more interaction between the two respective processes than the theory suggests. Researchers have argued that neither "fast" nor "slow" ToM can occur in isolation of the other [16]. This highlights the difficulty in focussing on either explicit or implicit elements of ToM, as mentioned earlier. The present paradigm and results demonstrates the variety of influences on ToM and the difficulties in untangling them.

There are significant differences in egocentrism vs altercentrism across groups, supporting hypotheses 1 (no cue) and 3 (full cue), but not hypothesis 2 (early cue). Whether the cue was shown throughout the video or only early-on, participants largely considered their own

knowledge rather than the other's. When there was no cue, the agent's knowledge was considered more (Figure 8). This suggests a high sensitivity to egocentric information in the context of ToM, regardless of when it is shown, consistent with previous literature [4]. The results suggest a preference for egocentrism over perspective-taking regardless of when an egocentric cue is available.

It is noteworthy that the explanations of participants' choices differed significantly across conditions while the decisions themselves did not differ significantly. However, as mentioned above, the quantitative comparison was not far from reaching the significance level.

7.3. Model

A computational model based on the data and observations presented above is currently in development. The model will conceptualise the switch between whose beliefs, actions, and goals are cognitively represented at all and/or drawn upon for reasoning about another's mental states. A key goal for the model is to explore ways to constrain the recursive aspects of ToM that require predictions of predictions (Figure 9). The deeper the level of ToM, the more possibilities

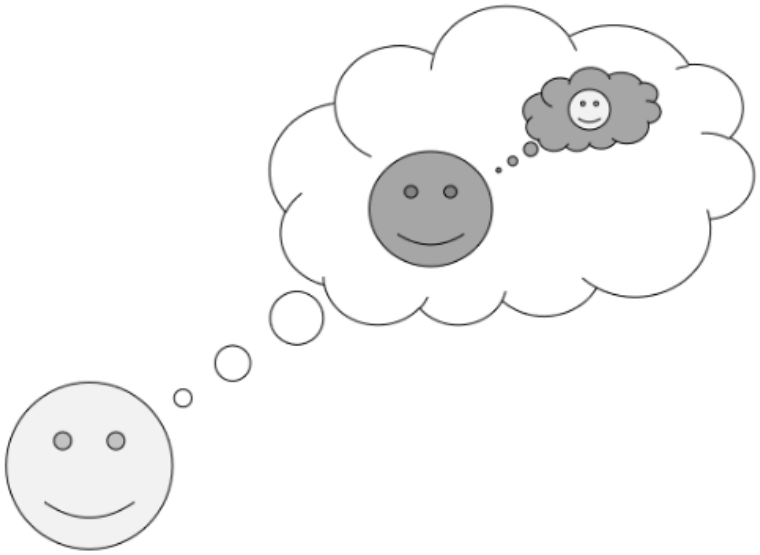


Figure 9: Recursion in ToM.

need to be considered when a prediction is generated (Figure 10). This computation is even more complex in a sequential decision-making setting over multiple steps in time. The present data suggests a bias towards using lower-order information even when next-order information is available. For example, even when it was possible for them to reason about the agent's beliefs or previous actions, many participants used their own tendencies or strategies to predict the agent's actions. Interestingly, there is also tendency by participants to make at least some prediction rather than indicating that they did not know where the agent would go.

A key implication for modelling ToM is therefore that the distinction between a fast and a slow system appears less fundamental than considering the information a human has previously

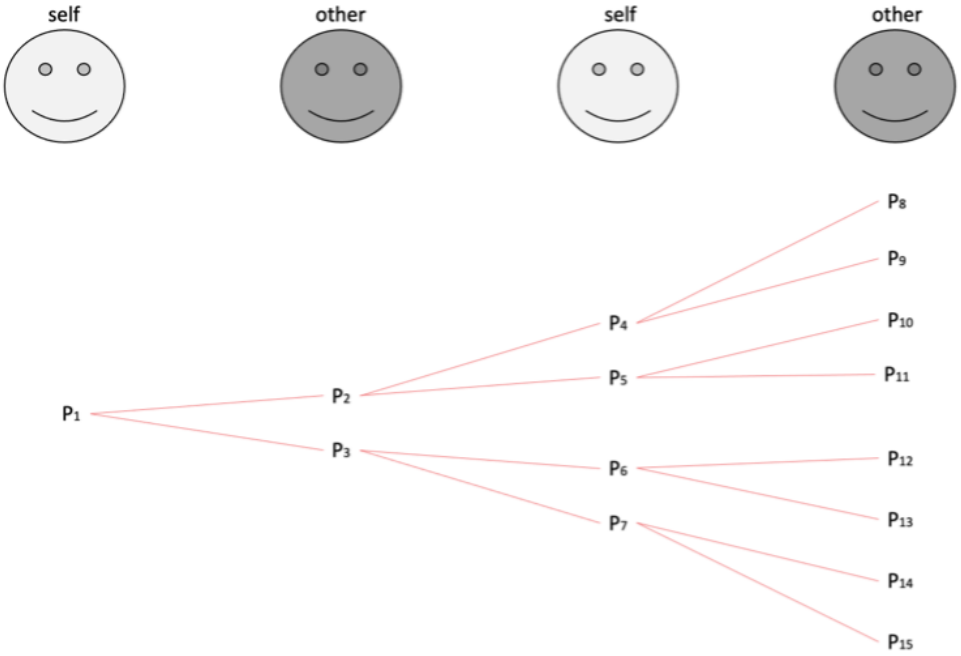


Figure 10: Possible predictions increase dramatically with higher-order ToM.

gathered. Participant responses greatly reflect their own prior experiences. This highlights the inherent relevance of subjective elements, such as priming effects or biases. Interestingly, these have not been applied in a majority of ToM models to date [20]

7.4. General discussion

ToM use largely varies across situations and individuals. Like with other aspects of cognition, humans develop strategies and habits which can differ considerably from one person to the next. For example, there are large differences in societal expectations across the world [28], and languages differ in the extent to which they include the vocabulary needed to communicate ToM concepts [29]. Evidence suggests that factors such as mood contribute to the extent of perspective-taking. For instance, uncertainty has been associated with more self-focussed reasoning [12].

All people are biased in their beliefs and perspectives, shaped by their own experiences. These biases are at the core of the very mechanism which makes human cognition so fast and efficient. But the speed and efficiency of human cognition doesn't mean that it is flawless. Biases come at the cost of errors which are often very difficult to detect [30]. This highlights an important question to consider as part of this research: Are the mechanisms required to develop robust, efficient, and dynamic artificial ToM worth the cost of the errors that are a fundamental part of human cognition?

On the other hand, considering the biases and flaws that are part of human ToM is fundamental to AI successfully and accurately predicting human behaviour. Furthermore, these insights may

assist current efforts to reduce biases and teach critical, reflected thinking in social contexts. Working beyond the biases in ToM may result in strategies helpful to both humans and artificial agents in understanding and managing social situations. Modelling human strengths and weaknesses is critical to realising AI powered assistive technology, particularly effective human-AI teamwork.

This study has indicated the extent to which 1) people can employ very different reasoning to come to the same conclusion and 2) the value of qualitative rather than quantitative research to identify these subtle differences. Rather than thinking about ToM as a binary ability, which is either present or absent, the present results show the possible variance in mental model complexity. We suggest that mental model depth and complexity is an important factor contributing to flexible and efficient ToM use. Future research will need to explore more deeply how mental models are structured, how or when the depth of mental models changes, and how exactly more slow, elaborate ToM differs from faster, biased ToM.

Acknowledgments

The authors are especially grateful to Jan Pöppel, Bielefeld University, Germany. This study builds on his research on Theory of Mind and his maze tool, ideas for designing the study, and general advice have been most helpful for collecting the data presented here.

References

[1] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, The behavioural and brain sciences 4 (1978) 515–526.

[2] M. Hecht, J. De Vito, L. Guerrero, Perspectives on nonverbal communication: codes, functions, and contexts, in: The Nonverbal Communication Reader, Waveland Press Long Grove, 1999, pp. 3–18.

[3] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B. J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z. L. Wang, R. Wood, The grand challenges of Science Robotics, SCIENCE ROBOTICS 3 (2018) 1–14.

[4] B. Keysar, S. Lin, D. J. Barr, Limits on theory of mind use in adults, Cognition 89 (2003) 25–41. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027703000647. doi:10.1016/S0010-0277(03)00064-7.

[5] C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing, Nature Human Behaviour 1 (2018) 1–10. Publisher: Nature Publishing Group.

[6] S. M. Schaafsma, D. W. Pfaff, R. P. Spunt, R. Adolphs, Deconstructing and reconstructing theory of mind, Trends in Cognitive Sciences 19 (2015) 65–72. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661314002502. doi:10.1016/j.tics.2014.11.007.

[7] F. Quesque, Y. Rossetti, What do theory-of-mind tasks actually measure? Theory and practice, Perspectives on Psychological Science 15 (2020) 384–396. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[8] C. Hughes, S. R. Jaffee, F. Happé, A. Taylor, A. Caspi, T. E. Moffitt, Origins of individual differences in theory of mind: From nature to nurture?, Child development 76 (2005) 356–370. Publisher: Wiley Online Library.

[9] C. Kobayashi, G. H. Glover, E. Temple, Cultural and linguistic effects on neural bases of 'Theory of Mind' in American and Japanese children, Brain Research 1164 (2007) 95–107. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006899307013418. doi:10.1016/j.brainres.2007.06.022.

[10] M. Freundlieb, M. Kovács, N. Sebanz, When do humans spontaneously adopt another's visuospatial perspective?, Journal of Experimental Psychology: Human Perception and Performance 42 (2015) 401–412. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000153. doi:10.1037/xhp0000153.

[11] D. Samson, I. A. Apperly, J. J. Braithwaite, B. J. Andrews, S. E. Bodley Scott, Seeing it their way: Evidence for rapid and involuntary computation of what other people see., Journal of Experimental Psychology: Human Perception and Performance 36 (2010) 1255–1266. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0018729. doi:10.1037/a0018729.

[12] A. R. Todd, M. Forstmann, P. Burgmer, A. W. Brooks, A. D. Galinsky, Anxious and egocentric: how specific emotions influence perspective taking, Journal of Experimental Psychology: General 144 (2015) 374–391.

[13] I. A. Apperly, S. A. Butterfill, Do humans have two systems to track beliefs and belief-like states?, Psychological Review 116 (2009) 953–970. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0016923. doi:10.1037/a0016923.

[14] S. Chaiken, Y. Trope, Dual-process theories in social psychology, Guilford Press, 1999.

[15] D. Kahneman, Thinking, fast and slow, Macmillan, 2011.

[16] A. Avramides, M. Parrott, Knowing other Minds, Oxford University Press, 2019.

[17] T. Burge, Do infants and nonhuman animals attribute mental states?, Psychological Review 125 (2018) 409–434. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/rev0000091. doi:10.1037/rev0000091.

[18] M. Freundlieb, M. Kovács, N. Sebanz, Reading Your Mind While You Are Reading—Evidence for Spontaneous Visuospatial Perspective Taking During a Semantic Categorization Task, Psychological Science 29 (2018) 614–622. URL: http://journals.sagepub.com/doi/10.1177/0956797617740973. doi:10.1177/0956797617740973.

[19] W. Xu, Toward human-centered AI: a perspective from human-computer interaction, interactions 26 (2019) 42–46. Publisher: ACM New York, NY, USA.

[20] S. V. Albrecht, P. Stone, Autonomous agents modelling other agents: A comprehensive survey and open problems, Artificial Intelligence 258 (2018) 66–95. Publisher: Elsevier.

[21] A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science, BEHAVIORAL AND BRAIN SCIENCES 36 (2013) 1–73.

[22] J. Perner, H. Wimmer, "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children, Journal of experimental child psychology 39 (1985) 437–471. Publisher: Elsevier.

[23] P. J. Gmytrasiewicz, E. H. Durfee, A Rigorous, Operational Formalization of Recursive Modeling., in: ICMAS, 1995, pp. 125–132.

[24] S. C. Marsella, D. V. Pynadath, S. J. Read, PsychSim: Agent-based modeling of social interactions and in□uence, Proceedings of the international conference on cognitive

modelling 36 (2004) 243–248.
[25] D. V. Pynadath, S. C. Marsella, PsychSim: Modeling theory of mind with decision-theoretic agents, in: IJCAI, volume 5, 2005, pp. 1181–1186.
[26] Y. Zeng, Y. Zhao, T. Zhang, D. Zhao, F. Zhao, E. Lu, A Brain-Inspired Model of Theory of Mind, Frontiers in Neurorobotics 14 (2020) 1–17. URL: https://www.frontiersin.org/article/10.3389/fnbot.2020.00060/full. doi:10.3389/fnbot.2020.00060.
[27] J. Pöppel, S. Marsella, S. Kopp, Less Egocentric Bias in Theory of Mind When Observing Agents in Unbalanced Decision Problems, Proceedings of CogSci, preprint (2021).
[28] M. J. Gelfand, J. L. Raver, L. Nishii, L. M. Leslie, J. Lun, B. C. Lim, L. Duan, A. Almaliach, S. Ang, J. Arnadottir, others, Differences between tight and loose cultures: A 33-nation study, science 332 (2011) 1100–1104. Publisher: American Association for the Advancement of Science.
[29] J. E. Pyers, A. Senghas, Language promotes false-belief understanding: Evidence from learners of a new sign language, Psychological science 20 (2009) 805–812. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
[30] T. Bartlett, Can we really measure implicit bias? Maybe not, The chronicle of higher education 63 (2017) B6–B7.